



Information Extraction and Aggregation from Unstructured Web Data for Business Profiling

Student Team	: Alexander Michels, Himanshu Ahuja, Liang Shi
Academic Mentor	: Shadi Shahsavari
Industry Mentor	: Dr. Stephen DeSalvo, Urjit Patel

Praedicat: An Insurance Tech Company

- Determine litigation risks
- Predict the likely amount of losses



1. Manual
Search

2. Credible
Database

3. Forward-
looking Models

4. Predict Likely
Losses

Where do we fit in?

RIPS Team
Automating



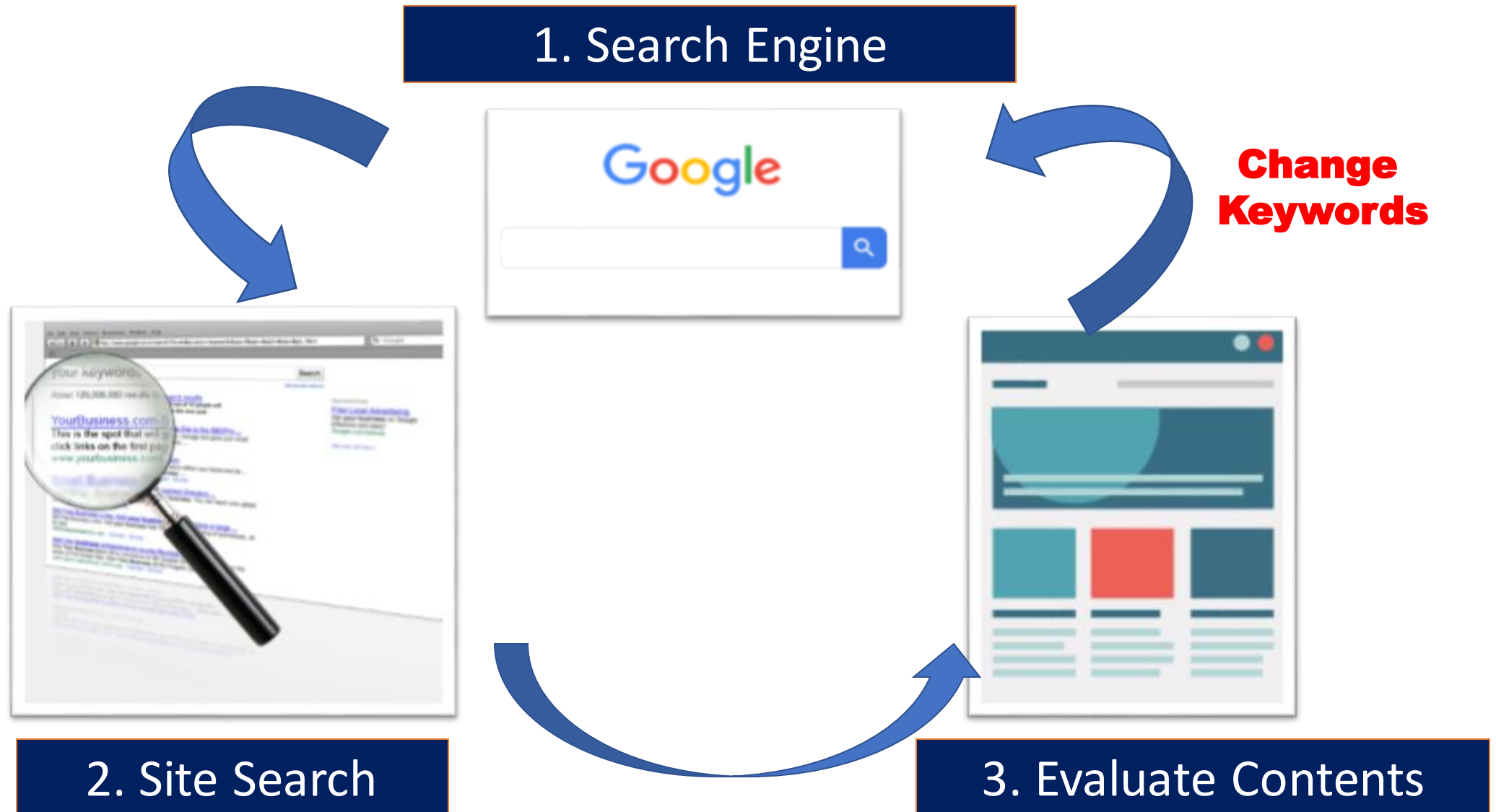
1. Manual
Search

2. Credible
Database

3. Forward-
looking Models

4. Predict Likely
Losses

Manual Search Process



Difficulty of Searching Information

Less Indicative of Litigation Risks

TRI Facility Report

Home Multisystem Search Topic Searches System Data Searches About the Data D

TRI Facility Report: 3M CO - DECATUR(35602MCMPNSTATE)
Facility Information

[FACILITY INFORMATION](#) [CHEMICALS](#) [POLLUTION PREVENTION \(P2\)](#) [WASTE MANAGEMENT](#) [REL](#)

Facility Name	3M CO - DECATUR	TRI ID	35602MCMPNSTATE
Address	1400 STATE DOCKS RD DECATUR, AL, 35601	FRS ID	110000367567
Mailing Name	3M CO - DECATUR	DUNS Number	006173082
Mailing Address	1400 STATE DOCKS RD DECATUR, AL, 35601	Parent Company	3M CO
County	MORGAN	Public Contact	MICHELLE HOWELL
EPA Region	4	Phone	(256) 552-6300
Latitude	34.641667	Tribal	NA
Longitude	-87.038611	BIA Tribal Code	NA
NAICS	325998 All Other Miscellaneous Chemical Product and Preparation Manufacturing	Industry Sector	325 Chemicals

- Government Databases

More Indicative of Litigation Risks

CNN Health • Food • Fitness • Wellness • Parenting • Vital Signs [Live TV](#) [U.S. Edition](#) [Search](#)

Jurors give \$289 million to a man they say got cancer from Monsanto's Roundup weedkiller

By [Holly Yan, CNN](#)
Updated 9:28 PM ET, Sat August 11, 2018



Judge reads final verdict in Monsanto case 01:32

[CNN] — San Francisco jurors just ruled that Roundup, the most popular weedkiller in the world, gave a former school groundskeeper terminal cancer.

More from CNN

- [Colorado man arrested after pregnant wife and two daughters go...](#)
- [Why Paul Manafort isn't wearing socks](#)

Pod Content [Fullbrain](#)



Wow - These New Luxury Cars Are Loaded

- News

Structured Web Pages

Facility Report for 3M

TRI Facility Report: 3M CO - DECATUR(35602MCPNSTATE)

Facility Information

FACILITY INFORMATION CHEMICALS POLLUTION PREVENTION (P2) WASTE MANAGEMENT REI			
Facility Name	3M CO - DECATUR	TRI ID	35602MCPNSTATE
Address	1400 STATE DOCKS RD DECATUR, AL, 35601	FRS ID	110000367567
Mailing Name	3M CO - DECATUR	DUNS Number	006173082
Mailing Address	1400 STATE DOCKS RD DECATUR, AL, 35601	Parent Company	3M CO
County	MORGAN	Public Contact	MICHELLE HOWELL
EPA Region	4	Phone	(256) 552-6300
Latitude	34.641667	Tribe	NA
Longitude	-87.038611	BIA Tribal Code	NA
NAIC(S)	325998 All Other Miscellaneous Chemical Product and Preparation Manufacturing	Industry Sector	325 Chemicals
Last Form	2017		

Facility Report for Samsung

TRI Facility Report: SAMSUNG AUSTIN SEMICONDUCTOR(78754SMSNG12100)

Facility Information

FACILITY INFORMATION CHEMICALS POLLUTION PREVENTION (P2) WASTE MANAGEMENT REI			
Facility Name	SAMSUNG AUSTIN SEMICONDUCTOR	TRI ID	78754SMSNG12100
Address	12100 SAMSUNG BLVD AUSTIN, TX, 78754	FRS ID	110000465531
Mailing Name	SAMSUNG AUSTIN SEMICONDUCTOR	DUNS Number	
Mailing Address	12100 SAMSUNG BLVD AUSTIN, TX, 78754	Parent Company	NA
County	TRAVIS	Public Contact	TIM JONES
EPA Region	6	Phone	(512) 799-8877
Latitude	30.375	Tribe	NA
Longitude	-97.631386	BIA Tribal Code	NA
NAIC(S)	334413 Semiconductor and Related Device Manufacturing	Industry Sector	334 Computers and Electronic Products
Last Form	2017		

Structured Web Pages

Facility Report for 3M

TRI Facility Report: 3M CO - DECATUR(35602MCPNSTATE)

Facility Information

FACILITY INFORMATION CHEMICALS POLLUTION PREVENTION (P2) WASTE MANAGEMENT RE			
Facility Name	3M CO - DECATUR	TRI ID	35602MCPNSTATE
Address	1400 STATE DOCKS RD DECATUR, AL, 35601	FRS ID	110000367567
Mailing Name	3M CO - DECATUR	DUNS Number	006173082
Mailing Address	1400 STATE DOCKS RD DECATUR, AL, 35601	Parent Company	3M CO
County	MORGAN	Public Contact	MICHELLE HOWELL
EPA Region	4	Phone	(256) 552-6300
Latitude	34.641667	Tribe	NA
Longitude	-87.038611	BIA Tribal Code	NA
NAIC(S)	325998 All Other Miscellaneous Chemical Product and Preparation Manufacturing	Industry Sector	325 Chemicals
Last Form	2017		

Facility Report for Samsung

TRI Facility Report: SAMSUNG AUSTIN SEMICONDUCTOR(78754SMSNG12100)

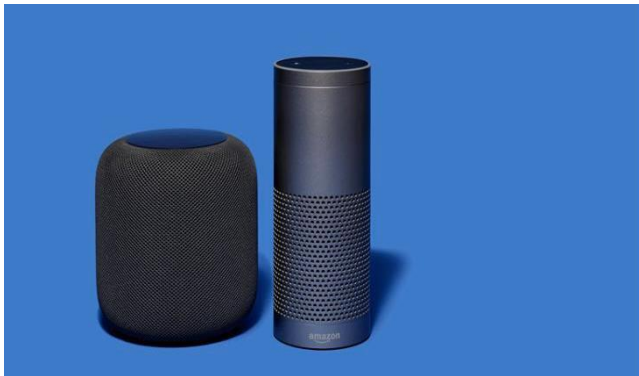
Facility Information

FACILITY INFORMATION CHEMICALS POLLUTION PREVENTION (P2) WASTE MANAGEMENT RE			
Facility Name	SAMSUNG AUSTIN SEMICONDUCTOR	TRI ID	78754SMSNG12100
Address	12100 SAMSUNG BLVD AUSTIN, TX, 78754	FRS ID	110000465531
Mailing Name	SAMSUNG AUSTIN SEMICONDUCTOR	DUNS Number	
Mailing Address	12100 SAMSUNG BLVD AUSTIN, TX, 78754	Parent Company	NA
County	TRAVIS	Public Contact	TIM JONES
EPA Region	6	Phone	(512) 799-8877
Latitude	30.375	Tribe	NA
Longitude	-97.631386	BIA Tribal Code	NA
NAIC(S)	334413 Semiconductor and Related Device Manufacturing	Industry Sector	334 Computers and Electronic Products
Last Form	2017		

Unstructured Web Pages

Alexa vs. Siri vs. Google: Which Can Carry on a Conversation Best?

By KEITH COLLINS and CADE METZ AUG. 17, 2018



▶ Who won the Giants game last night?

AMAZON ECHO

On December 31st, the Giants beat the Redskins 18 to 10. They'll play on August 9th at 7 p.m. at home against the Browns.

Elon Musk Details 'Excruciating' Personal Toll of Tesla Turmoil



Elon Musk, the chairman and chief executive of the electric-car maker Tesla. "This past year has been the most difficult and painful year of my career," he said. Noah Berger/Bloomberg

By David Gelles, James B. Stewart,
Jessica Silver-Greenberg and Kate Kelly

Give Your Old Computer New Life

If you're not ready to buy a whole new system, you might be able to add new parts and upgrade your aging machine for less than a few hundred dollars.



By J. D. Biersdorfer

Aug. 17, 2018

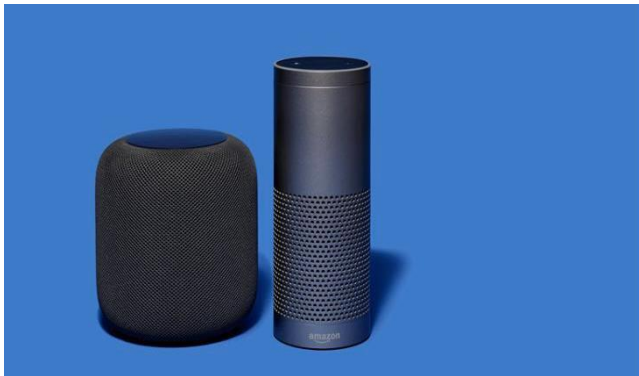


Q. *My computer is old, and defragmenting the drive is not going to make it any faster. I can add memory chips, but what else can I do to speed things up without spending a ton or buying a new machine?*

Unstructured Web Pages

Alexa vs. Siri vs. Google: Which Can Carry on a Conversation Best?

By KEITH COLLINS and CADE METZ AUG. 17, 2018



▶ Who won the Giants game last night?

AMAZON ECHO

On December 31st, the Giants beat the Redskins 18 to 10. They'll play on August 9th at 7 p.m. at home against the Browns.

Elon Musk Details 'Excruciating' Personal Toll of Tesla Turmoil



Elon Musk, the chairman and chief executive of the electric-car maker Tesla. "This past year has been the most difficult and painful year of my career," he said. Noah Berger/Bloomberg

By David Gelles, James B. Stewart,
Jessica Silver-Greenberg and Kate Kelly

Give Your Old Computer New Life

If you're not ready to buy a whole new system, you might be able to add new parts and upgrade your aging machine for less than a few hundred dollars.



By J. D. Biersdorfer

Aug. 17, 2018



Q. *My computer is old, and defragmenting the drive is not going to make it any faster. I can add memory chips, but what else can I do to speed things up without spending a ton or buying a new machine?*

Problem Statement

How to automate information extraction,
classification, and fact-checking for unstructured
data on the Internet

Solution Overview

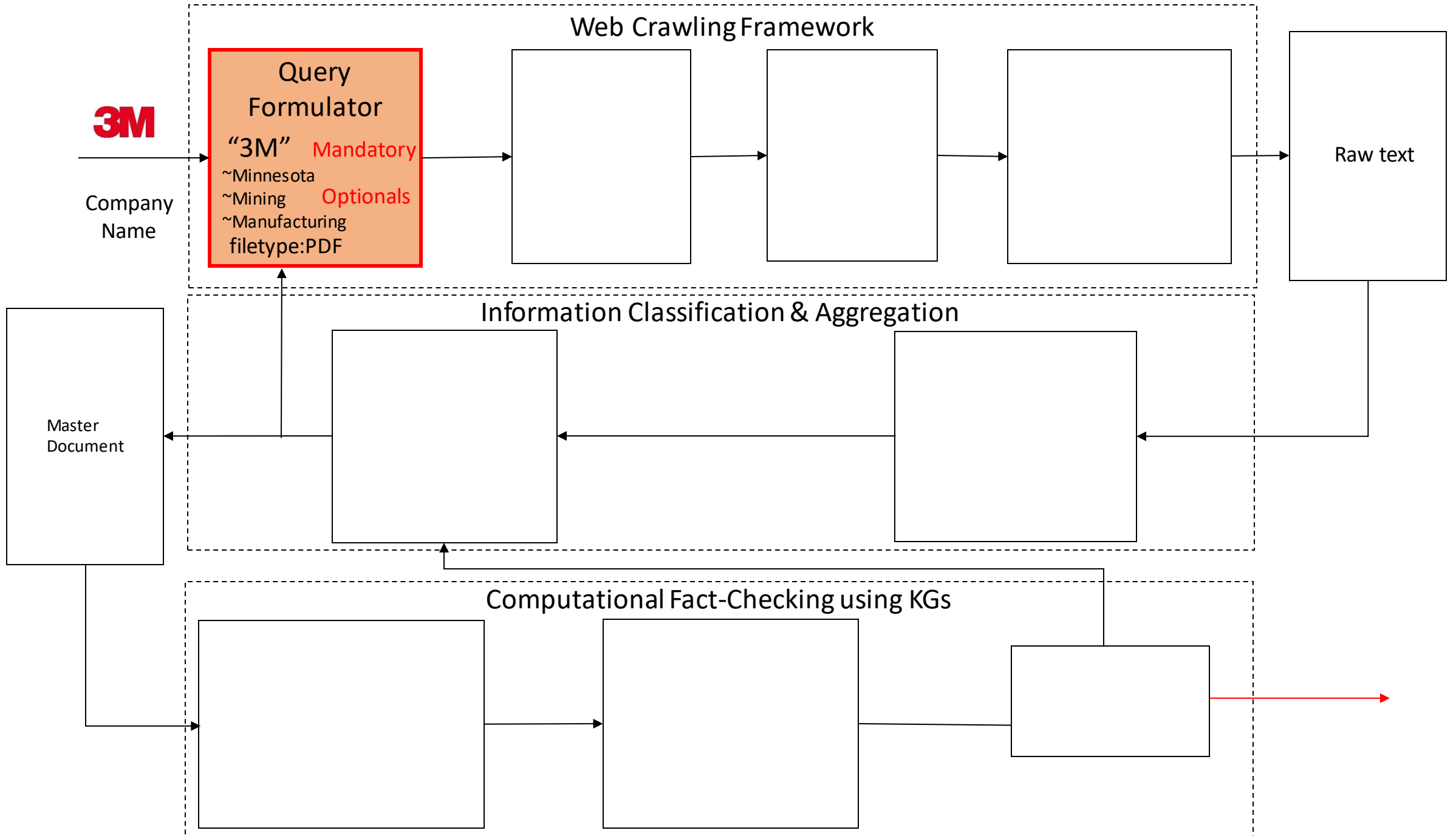
Web Crawling Framework



Information Classification & Aggregation



Computational Fact-Checking using KGs



Query Formulator: Asking about the right things!

Apple

Apple - Wikipedia
<https://en.wikipedia.org/wiki/Apple> ▼
An apple is a sweet, edible fruit produced by an apple tree (*Malus pumila*). Apple trees are cultivated worldwide, and are the most widely grown species in the ...
Family: Rosaceae Genus: *Malus*
Species: *M. pumila* Kingdom: Plantae
Apple Inc. · Fuji (apple) · Cooking apple · Fruit tree

People also ask

- What are the benefits of apples? ▼
- Which is the healthiest fruit? ▼
- What does Apple fruit symbolize? ▼
- Do apples help you lose weight? ▼

Feedback

Apple fruit nutrition facts and health benefits - Nutrition and You
<https://www.nutrition-and-you.com/apple-fruit.html> ▼
Delicious and crunchy apple fruit is one of the popular table fruits containing an impressive list of antioxidants and essential nutrients required for good health.

Apple
Fruit

An apple is a sweet, edible fruit produced by an apple tree. Apple trees are cultivated worldwide, and are the most widely grown species in the genus *Malus*. Wikipedia

Nutrition Facts
Apple ▼

Amount Per 1 medium (3" dia) (182 g)	
Calories 95	
	% Daily Value*
Total Fat 0.3 g	0%

Apple Inc

Apple
<https://www.apple.com/> ▼
Discover the innovative world of Apple and shop everything iPhone, iPad, Apple Watch, Mac, and Apple TV, plus explore accessories, entertainment, and expert ...

Search apple.com

Mac
MacBook Pro · MacBook · MacBook Air · Compare · iMac

iPhone
Explore iPhone, the world's most powerful personal device ...

Apple Support
Apple support is here to help. Learn more about popular ...

iTunes Store
Download iTunes · Music · Video · iTunes Charts · ...

Find a Store
Find a store. Complete store list. Do more of what you love ...

Accessories
Shop Apple accessories for Apple Watch, iPhone, iPad, iPod, and ...

Apple
Technology company

apple.com

Apple Inc. is an American multinational technology company headquartered in Cupertino, California, that designs, develops, and sells consumer electronics, computer software, and online services. Wikipedia

Stock price: AAPL (NASDAQ) \$191.61 +0.17 (+0.09%)
Jul 23, 4:00 PM EDT - Disclaimer

Founded: April 1, 1976, Cupertino, CA
Headquarters: Cupertino, CA

Zero useful results

'Apple Inc.' returns the right results.

PDF result mentions Rentokil Initial PLC involvement in window cleaning.

rentokil initial plc window cleaning

About 25,200 results (0.51 seconds)

Rentokil Initial plc
<https://www.rentokil-initial.com/> ▼
Leading in Pest Control. Our engine for growth ... Rentokil Initial plc welcomes the Commonwealth Heads of State commitment to halve the... 23 Apr 2018.
Missing: window | Must include: window

Protect & Enhance – Rentokil Initial plc
<https://www.rentokil-initial.com/our-services/other-services.aspx> ▼
Rentokil Specialist Hygiene is one of the leading providers of specialised deep cleaning and specialist industrial cleaning and disinfection services.
Missing: window | Must include: window

rentokil ~initial ~plc "window cleaning"

About 2,450 results (0.43 seconds)

Initial Facilities Management Ltd.: Private Company Information ...
<https://www.bloomberg.com/research/stocks/private/snapshot.asp?privcapld...> ▼
Initial Facilities Management Ltd. company research & investing information. ... deep cleaning, IT sanitizing, and window cleaning; clinical and dental waste ...

[PDF] Rentokil CR 09_AW.indd - Rentokil Initial plc
<https://www.rentokil-initial.com/~media/Files/R/Rentokil/.../ri-2009-cr-report.pdf> ▼
Rentokil Initial plc operates in over 50 countries across the world's major economic harnesses and all colleagues in the window cleaning businesses have ...

Query Formulator: How did we ask the right things?

The diagram illustrates two search queries and the components that make them effective. Red circles highlight specific parts of the queries, and red arrows point from these circles to explanatory text boxes.

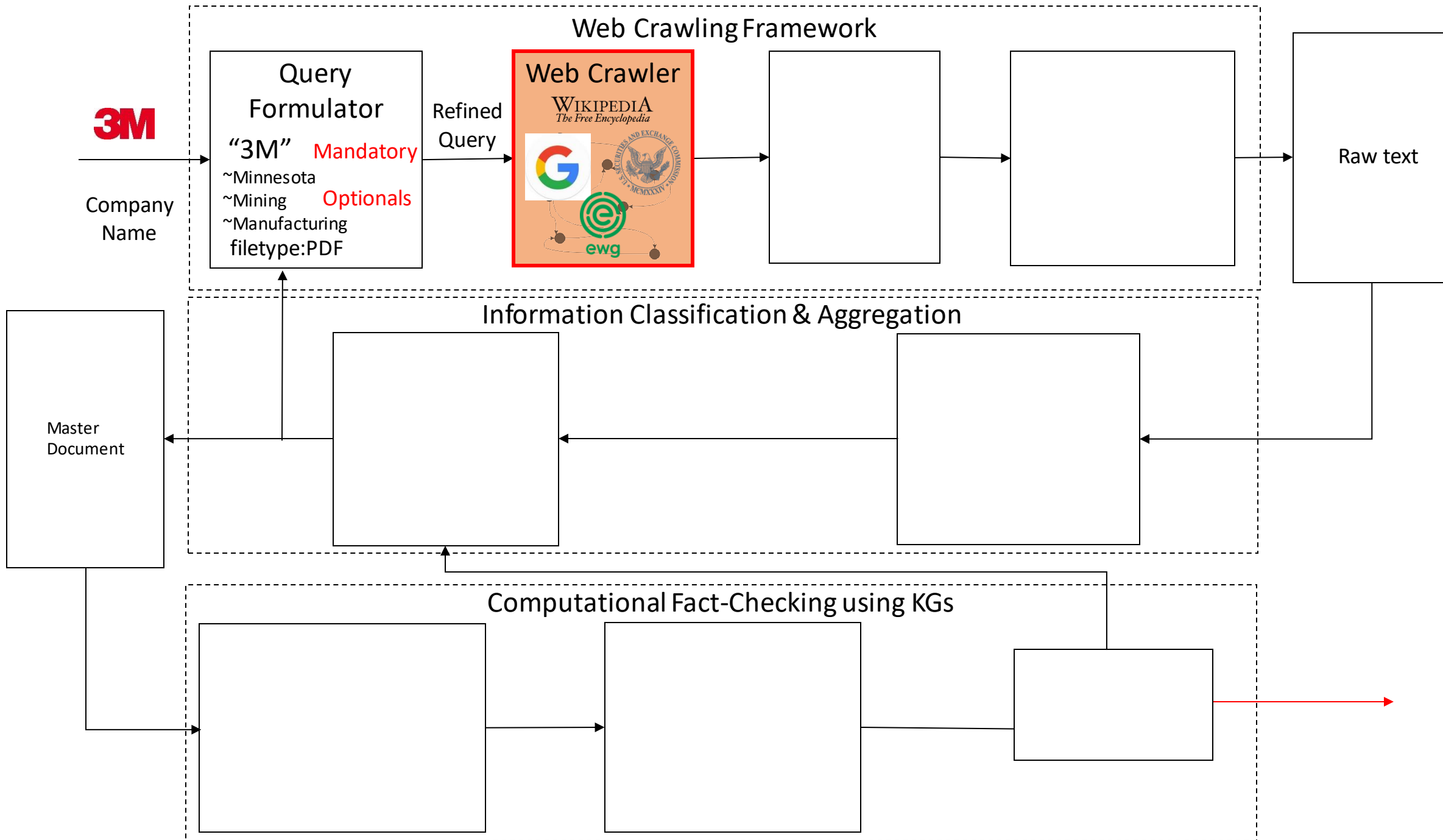
Left Query: `filetype:PDF apple ~inc ~computers "aluminium"`

- filetype:PDF** (circled) → **Mention the file-type**
- ~inc ~computers** (circled) → **Making some words optional**
- "aluminium"** (circled) → **Making keywords mandatory**

Right Query: `filetype:PDF minnesota mining ~and manufacturing ~3m "ethylene glycol"`

- minnesota mining ~and manufacturing ~3m** (circled) → **Name of the company**
- "ethylene glycol"** (circled) → **Optional alias**

Both queries show search results for PDF files. The left query results include "Supplier List - Apple" and "Supplier List 2015 - Apple". The right query results include "MATERIAL SAFETY 3M DATA SHEET 3M Center St. Paul, Minnesota ...".



Web Crawling: What is web Crawling?

rentokil initial PLC window cleaning

AllShoppingNewsMapsImagesMoreSettingsTools

About 25,200 results (0.51 seconds)

Rentokil Initial plc

<https://www.rentokil-initial.com/> ▼

Leading in Pest Control. Our engine for growth ... Rentokil Initial plc welcomes the Commonwealth Heads of State commitment to halve the... 23 Apr 2018.

Missing: window | Must include: [window](#)

Protect & Enhance – Rentokil Initial plc

<https://www.rentokil-initial.com/our-services/other-services.aspx> ▼

Rentokil Specialist Hygiene is one of the leading providers of specialised deep cleaning and specialist industrial cleaning and disinfection services.

Missing: window | Must include: [window](#)

Hygiene – Rentokil Initial plc

<https://www.rentokil-initial.com/our-services/hygiene.aspx> ▼

Rentokil Initial has undertaken a series of trials across Europe involving 100,000 people which monitored hand washing compliance. From low levels (as low as ...

Missing: window | Must include: [window](#)

Rentokil Specialist Hygiene: Specialist Cleaning and Disinfection ...

<https://www.rentokil-hygiene.co.uk/> ▼

Rentokil Specialist Hygiene provide specialist cleaning, kitchen deep cleaning and disinfection services for challenging environments.

Rentokil's pest control expertise for facility management | Rentokil

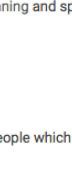
<https://www.rentokil.com/sector-insights/facilities-management/expertise/> ▼

Rentokil offers facility management companies market-leading expertise in pest ... fouling window ledges, balconies and pavements, spreading diseases and ... Industrial cleaning; Specialist disinfection; Washroom deep cleaning ... Initial Medical (visit Initial to find out more). 2018 Rentokil Initial plc Legal statement.

Initial - Experts in Washroom Hygiene, Floor Mats & Healthcare Waste

<https://www.initial.ie/> ▼

Initial Ireland, the leading provider of washroom hygiene service, medical waste management, floor mats, washroom supplies, cleaning service, toilet hygiene. ... 2018 Rentokil Initial plc and subject to the conditions in the legal statement.





Start

End

Initial Facilities Services						
£ million	Third Quarter			Year to Date		
	2010	2009	change	2010	2009	change
At 2009 constant exchange rates:						
Revenue	140.3	131.3	6.9%	407.8	411.3	(0.9%)
Adjusted operating profit (before one-off items and amortisation & impairment of intangible assets ¹)	6.8	5.7	19.3%	16.7	13.4	24.6%
At actual exchange rates:						
Adjusted operating profit (before one-off items and amortisation & impairment of intangible assets ¹)	6.7	5.7	17.5%	16.6	13.4	23.9%
¹ Other than computer software						

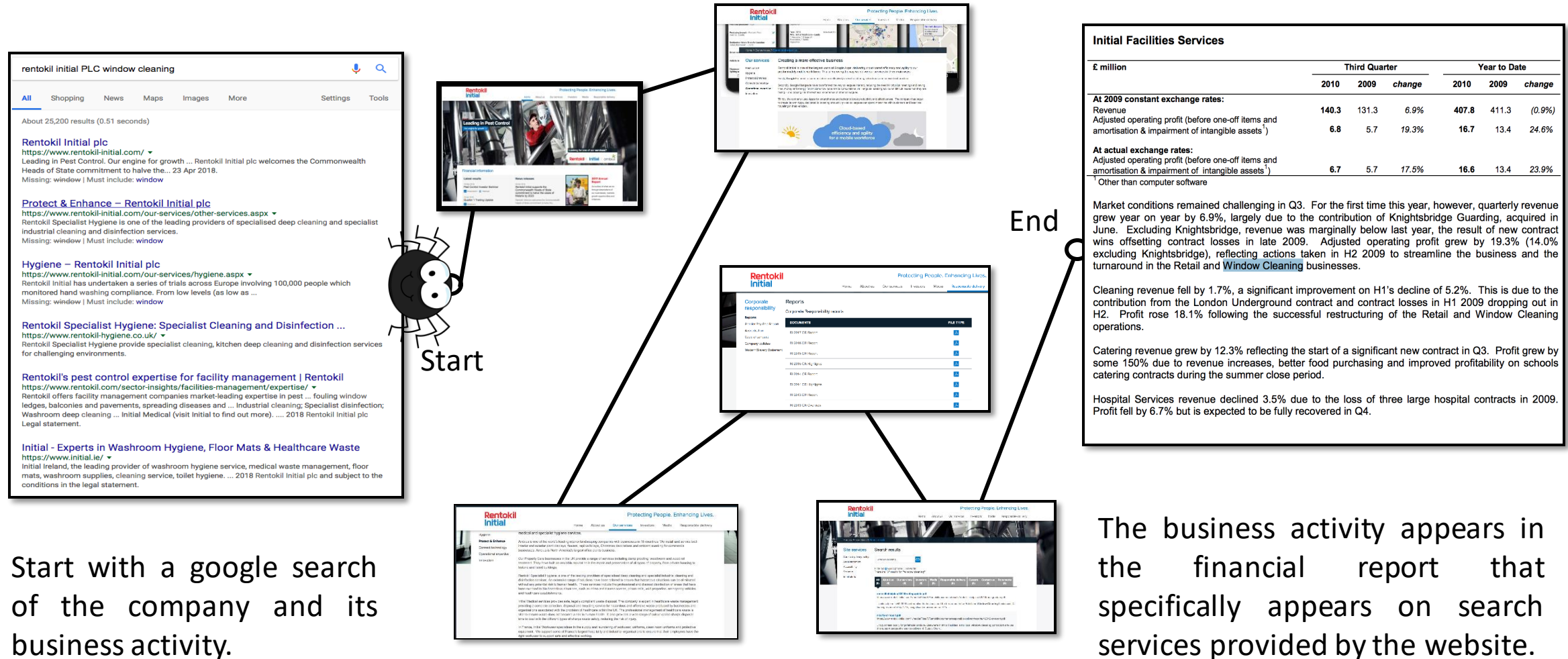
Market conditions remained challenging in Q3. For the first time this year, however, quarterly revenue grew year on year by 6.9%, largely due to the contribution of Knightsbridge Guarding, acquired in June. Excluding Knightsbridge, revenue was marginally below last year, the result of new contract wins offsetting contract losses in late 2009. Adjusted operating profit grew by 19.3% (14.0% excluding Knightsbridge), reflecting actions taken in H2 2009 to streamline the business and the turnaround in the Retail and Window Cleaning businesses.

Cleaning revenue fell by 1.7%, a significant improvement on H1's decline of 5.2%. This is due to the contribution from the London Underground contract and contract losses in H1 2009 dropping out in H2. Profit rose 18.1% following the successful restructuring of the Retail and Window Cleaning operations.

Catering revenue grew by 12.3% reflecting the start of a significant new contract in Q3. Profit grew by some 150% due to revenue increases, better food purchasing and improved profitability on schools catering contracts during the summer close period.

Hospital Services revenue declined 3.5% due to the loss of three large hospital contracts in 2009. Profit fell by 6.7% but is expected to be fully recovered in Q4.

Web Crawling: Unsupervised machines cannot be trusted





Web Crawling: Where and how far?

The problem:



We don't know how far to dig, and where to dig?

We don't know the credible sources and where the information lies on the credible sources.

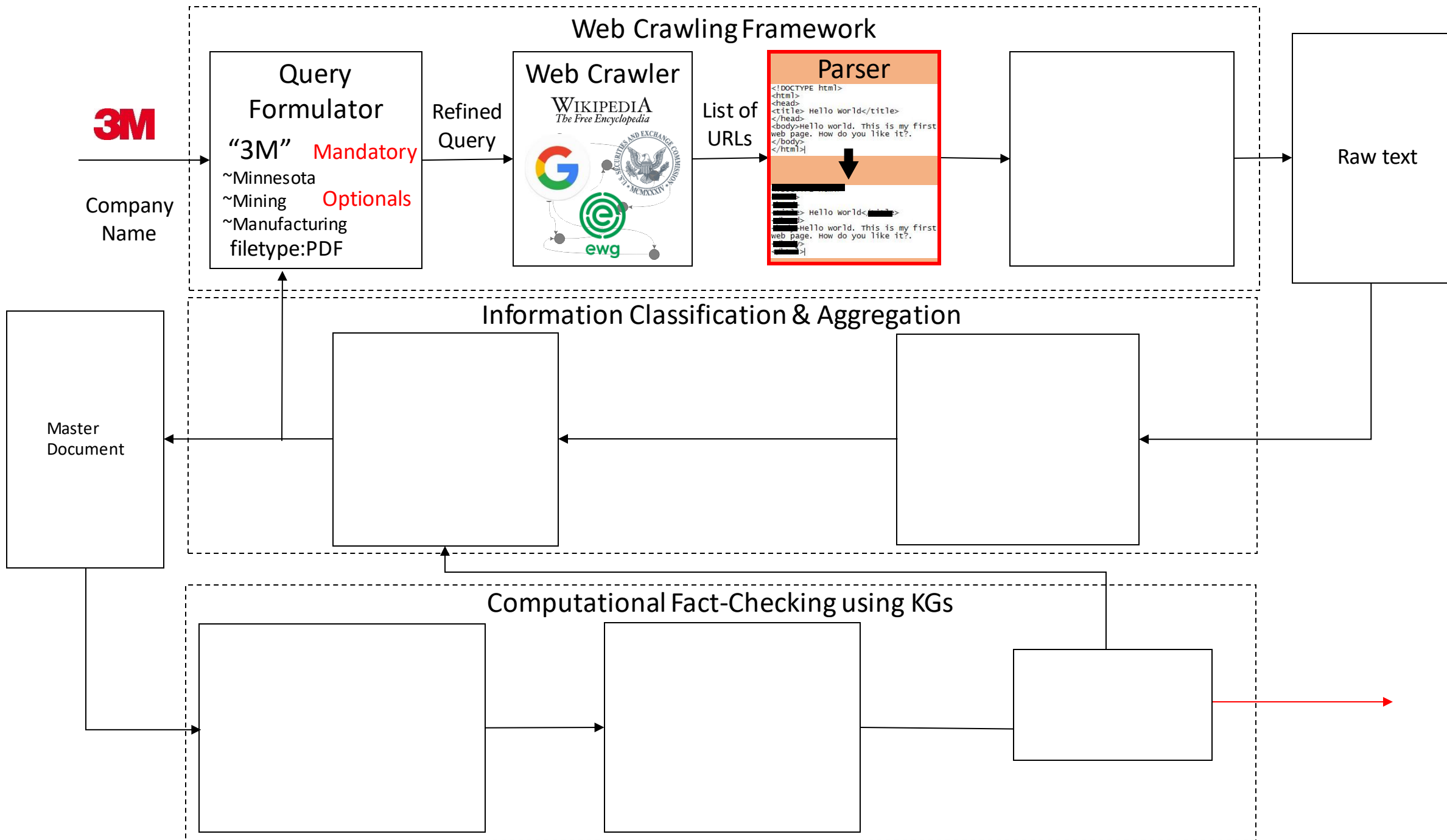
Web Crawling: Credible data to the rescue

DOW JONES & CO INC CIK#: 0000029924 (see all company filings)		
SIC: 2711 - NEWSPAPERS: PUBLISHING OR PUBLISHING & PRINTING		
State location: NY State of Inc.: DE Fiscal Year End: 1231		
(Assistant Director Office: 5)		
Get insider transactions for this issuer.		
Filter Results:	Filing Type:	Prior to: (YYYYMMDD)
Items 1 - 100 RSS Feed		
Filings	Format	Description
SC 13G/A	Documents	[Amend] Statement of acquisition of beneficial ownership b Acc-no: 0000070858-08-000217 (34 Act) Size: 11 KB
SC 13G/A	Documents	[Amend] Statement of acquisition of beneficial ownership b Acc-no: 0000070858-08-000126 (34 Act) Size: 19 KB
15-15D	Documents	Suspension of duty to report [Section 13 and 15(d)] Acc-no: 0001193125-07-268821 (34 Act) Size: 15 KB
4/A	Documents	[Amend] Statement of changes in beneficial ownership of s Acc-no: 0001140361-07-024624 Size: 23 KB
4/A	Documents	[Amend] Statement of changes in beneficial ownership of s Acc-no: 0001140361-07-024610 Size: 9 KB
15-12B	Documents	Securities registration termination [Section 12(b)] Acc-no: 0000895345-07-000712 (34 Act) Size: 22 KB
4	Documents	Statement of changes in beneficial ownership of securities Acc-no: 0001221662-07-000061 Size: 5 KB
25-NSE	Documents	Notification filed by national security exchange to report the Acc-no: 0000876661-07-000947 (34 Act) Size: 4 KB
4	Documents	Statement of changes in beneficial ownership of securities Acc-no: 0001140361-07-024430 Size: 6 KB
4	Documents	Statement of changes in beneficial ownership of securities Acc-no: 0001140361-07-024427 Size: 9 KB
4	Documents	Statement of changes in beneficial ownership of securities Acc-no: 0001221662-07-000060 Size: 8 KB
4	Documents	Statement of changes in beneficial ownership of securities Acc-no: 0000895345-07-000708 Size: 40 KB
4	Documents	Statement of changes in beneficial ownership of securities Acc-no: 0000895345-07-000707 Size: 38 KB
4	Documents	Statement of changes in beneficial ownership of securities Acc-no: 0000895345-07-000706 Size: 23 KB
4	Documents	Statement of changes in beneficial ownership of securities Acc-no: 0000895345-07-000705 Size: 11 KB
4	Documents	Statement of changes in beneficial ownership of securities Acc-no: 0000895345-07-000704 Size: 5 KB
4	Documents	Statement of changes in beneficial ownership of securities Acc-no: 0000895345-07-000703 Size: 16 KB
4	Documents	Statement of changes in beneficial ownership of securities Acc-no: 0000895345-07-000702 Size: 28 KB
4	Documents	Statement of changes in beneficial ownership of securities Acc-no: 0000895345-07-000701 Size: 16 KB

Dow Jones & Company Inc.	
	
DOW JONES	
Type	Subsidiary of News Corp.
Industry	News and Publishing
Founded	November 1882; 135 years ago 15 Wall Street, New York City, New York, U.S.
Founder	Charles Dow, Edward Jones, Charles Bergstresser
Headquarters	1211 Avenue of the Americas New York, NY 10036 U.S.
Key people	William Lewis (CEO) ^[1] ^[2]
Products	<i>The Wall Street Journal</i> <i>Barron's</i> Dow Jones Newswires Dow Jones Financial Information Services DJX Factiva MarketWatch (See complete products listing.)
Revenue	▲\$1.5 billion USD (2009)
Net income	▲\$386.56 million USD (2009)
Parent	News Corp
Website	dowjones.com 

Row #	Chemical
	 
1	1,1,1,2-TETRACHLORO-2-FLUOROETHANE
2	1,1,1,2-TETRACHLOROETHANE
3	1,1,1-TRICHLOROETHANE
4	1,1,2,2-TETRACHLOROETHANE
5	1,1,2-TRICHLOROETHANE
6	1,1-DICHLORO-1-FLUOROETHANE
7	1,1-DIMETHYL HYDRAZINE
8	1,2,3-TRICHLOROPROPANE
9	1,2,4-TRICHLOROBENZENE
10	1,2,4-TRIMETHYL BENZENE
11	1,2-BUTYLENE OXIDE
12	1,2-DIBROMO-3-CHLOROPROPANE
13	1,2-DIBROMOETHANE
14	1,2-DICHLORO-1,1,2-TRIFLUOROETHANE
15	1,2-DICHLORO-1,1-DIFLUOROETHANE
16	1,2-DICHLOROBENZENE
17	1,2-DICHLOROETHANE
18	1,2-DICHLOROETHYLENE
19	1,2-DICHLOROPROPANE
20	1,2-DIPHENYLHYDRAZINE

- Interestingly, the **structured data** (available on Federal websites & Wikipedia) is **also credible!**
- Designed specific crawlers to get data from specific databases.
- Created a baseline data to support unsupervised web crawling.



Parser:

Getting unstructured data

Use of text abundance to locate meaningful paragraphs.

Filtering out tags containing social media redirects.

Removing graphic contents, advertisements.



This undated photo shows 3M in St. Paul, Minn.

RELATED

- **Lawsuits against Wolverine claim 3 deaths tied to PFAS**
- **Judge: 3M should also answer to local PFAS lawsuits**
- **Plainfield Twp: New filters clearing out most PFAS**

GRAND RAPIDS, Mich. (WOOD) — Attorneys suing over Wolverine Worldwide's PFAS contamination describe a high-stakes game between the Rockford shoemaker and the company that made the chemical.

The **original lawsuits** filed by homeowners in northern Kent County named only Wolverine as a defendant, but amended complaints filed Monday now include 3M.

The lawsuits claim that 3M, the chemical company based in Minnesota, and the Rockford-based Wolverine each spent years separately covering up the impacts of the likely carcinogen.

Toxic Tap Water: Full Coverage

Click/tap here for what you need to know now.
[Read More »](#)



Trending Stories

Day care owner caught overbilling state by \$178K

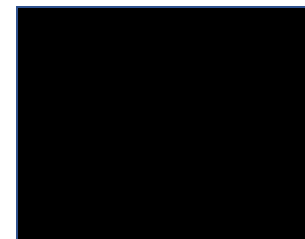
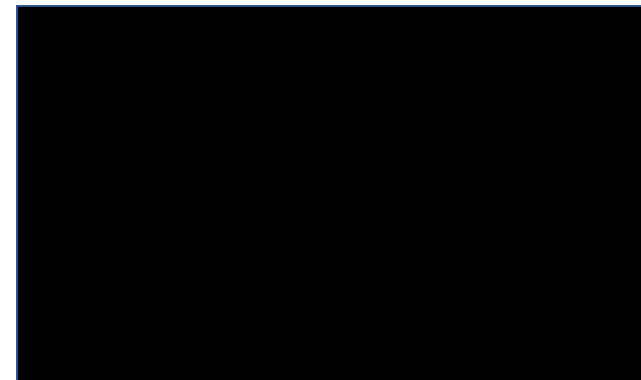
Source: 1 dead in 3-car crash on US-131

Storm Team 8 Forecast

Medical marijuana facility to bring 400 jobs to Marshall

Rockford couple with 14 sons get Lifetime movie

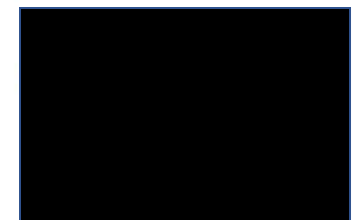
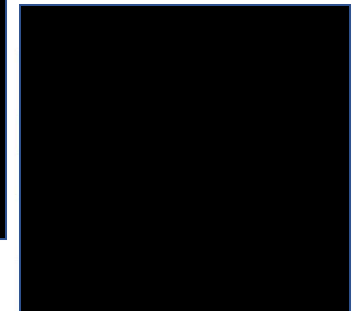
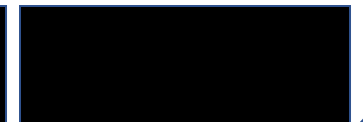
Photo Galleries

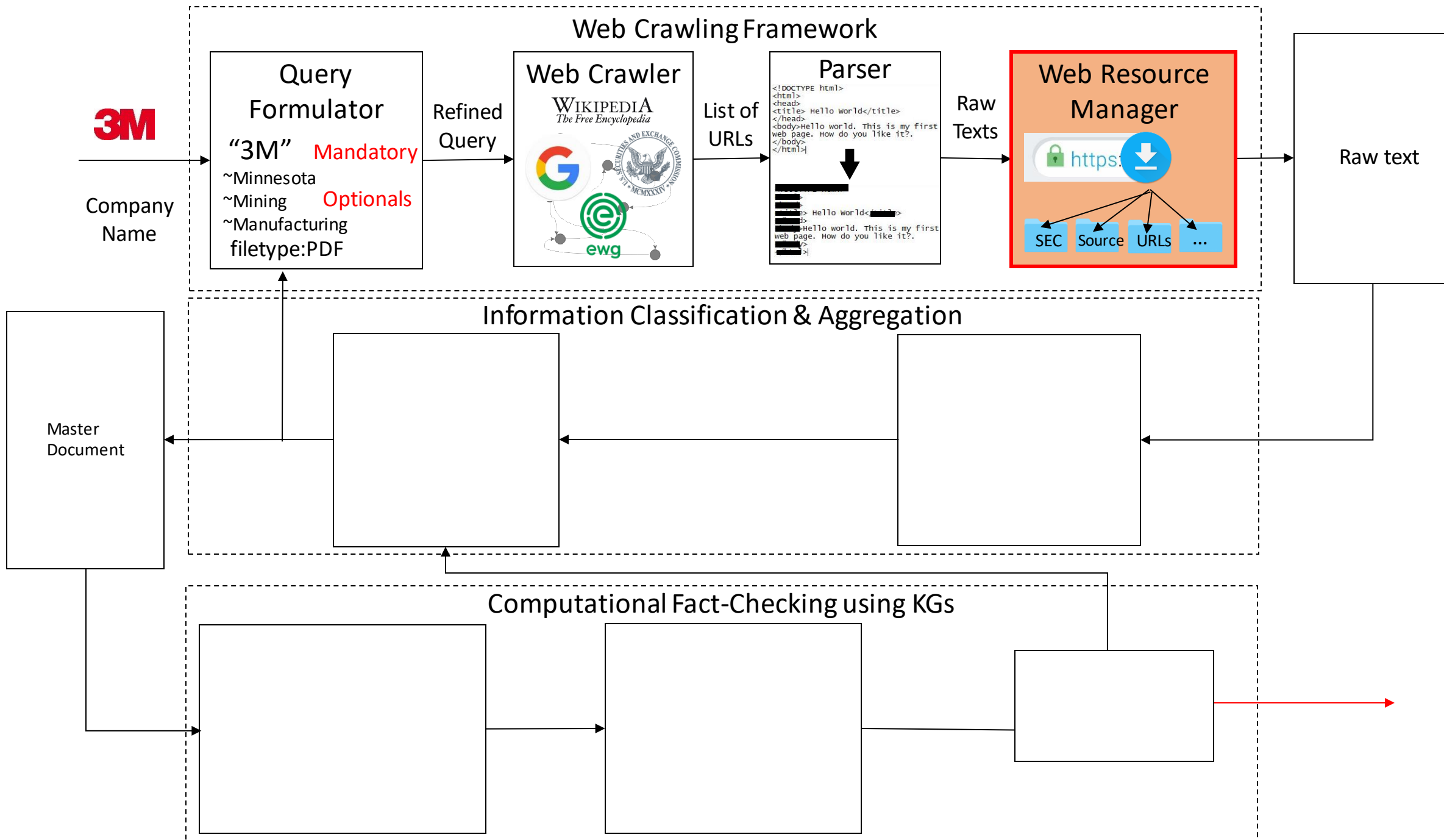


GRAND RAPIDS, Mich. (WOOD) — Attorneys suing over Wolverine Worldwide's PFAS contamination describe a high-stakes game between the Rockford shoemaker and the company that made the chemical.

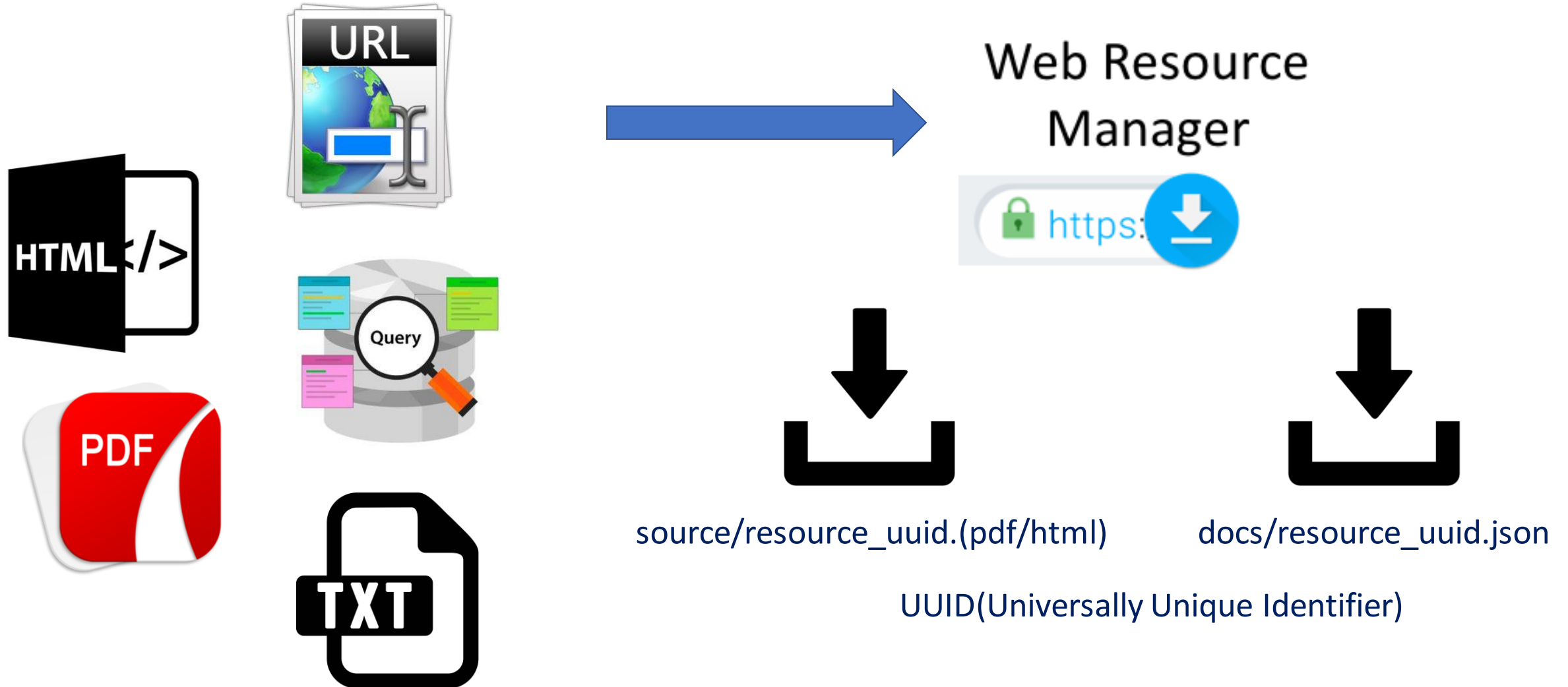
The **original lawsuits** filed by homeowners in northern Kent County named only Wolverine as a defendant, but amended complaints filed Monday now include 3M.

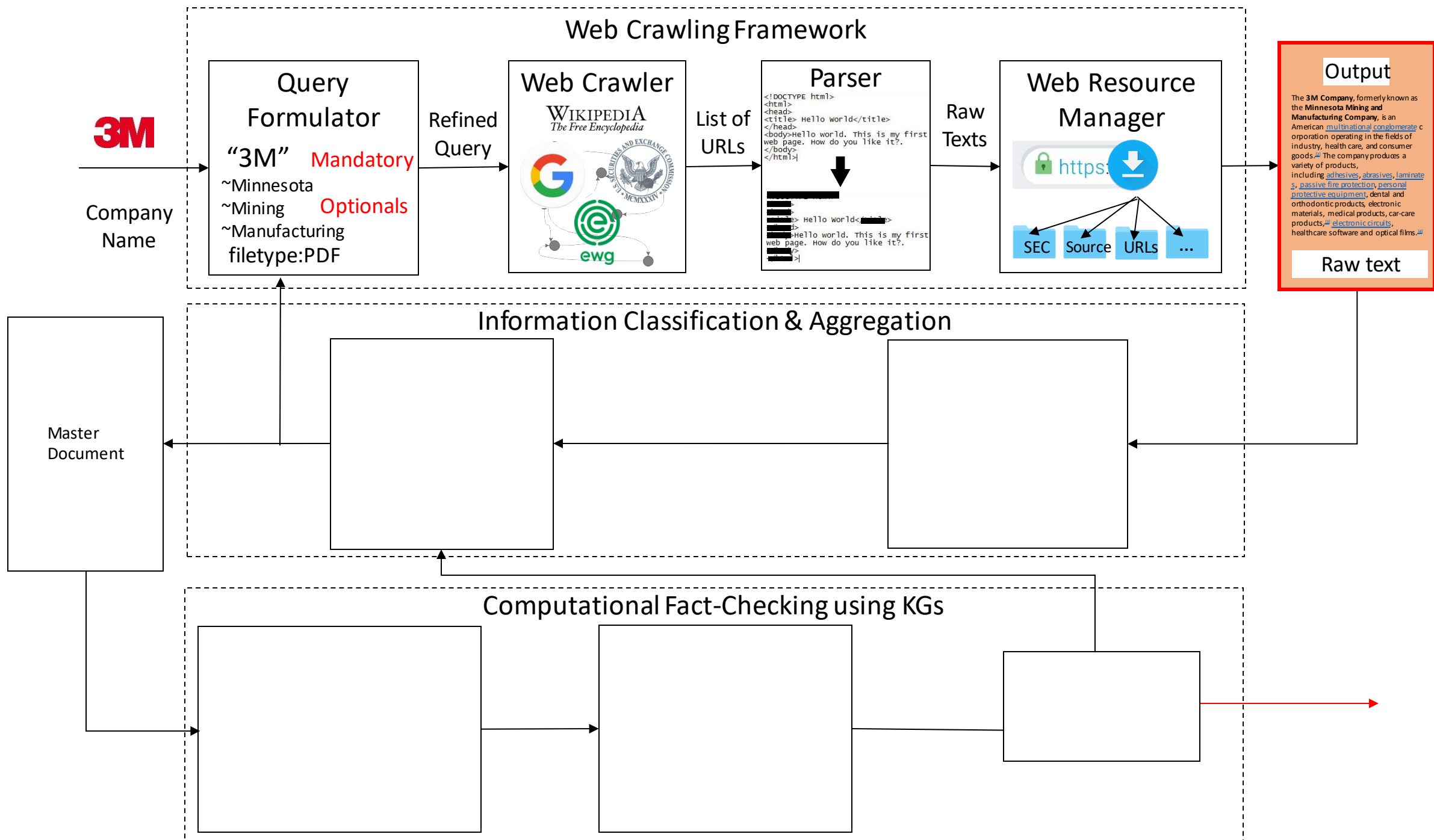
The lawsuits claim that 3M, the chemical company based in Minnesota, and the Rockford-based Wolverine each spent years separately covering up the impacts of the likely carcinogen.





Web Resource Manager:

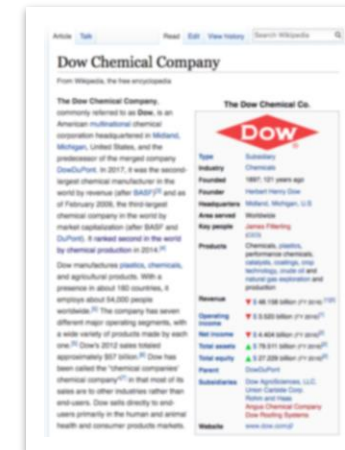
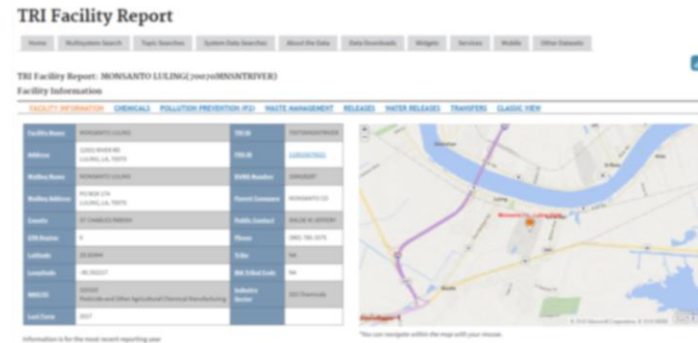
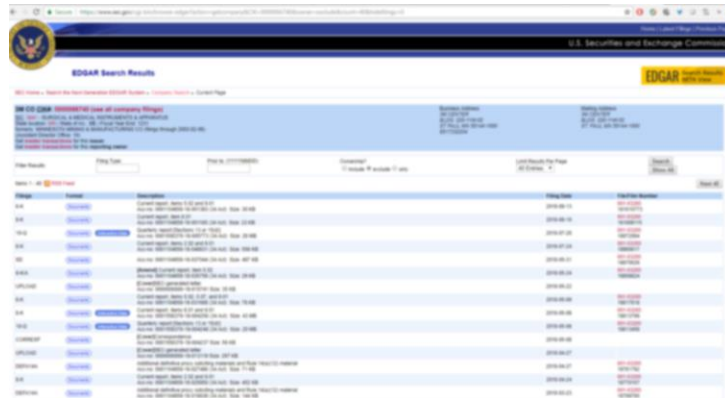


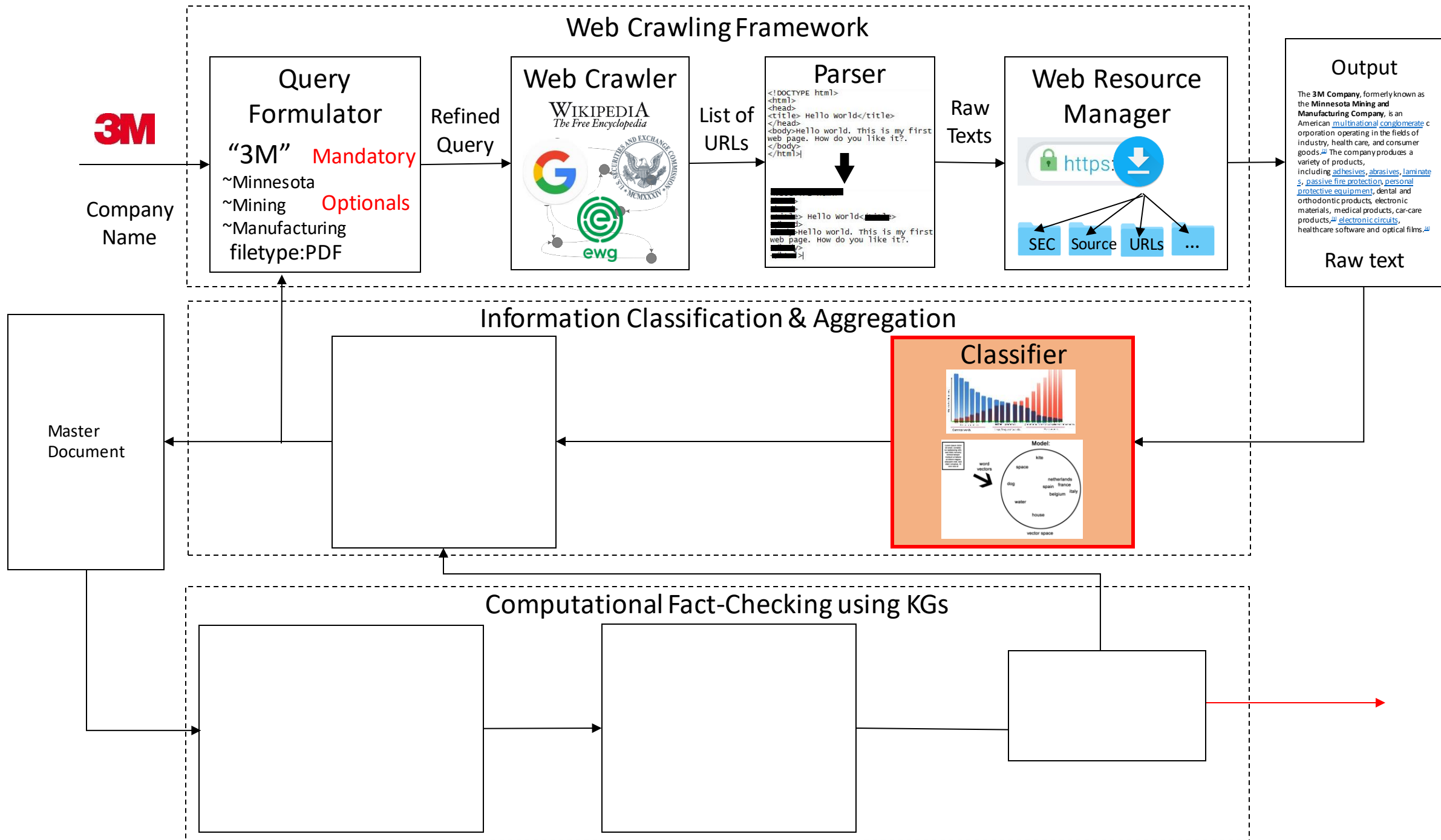


Outputs of Site Crawlers

Data

- Financial statements for 52,629 companies
- 21,202 Facility Reports
- Product and ingredient list for 4,535 companies
- Thousands of subsidiary structures
- Tens of thousands of Wikipedia pages





Self-Supervised Learning



Label

Labels its own
training examples
using heuristics

Train
Classifier

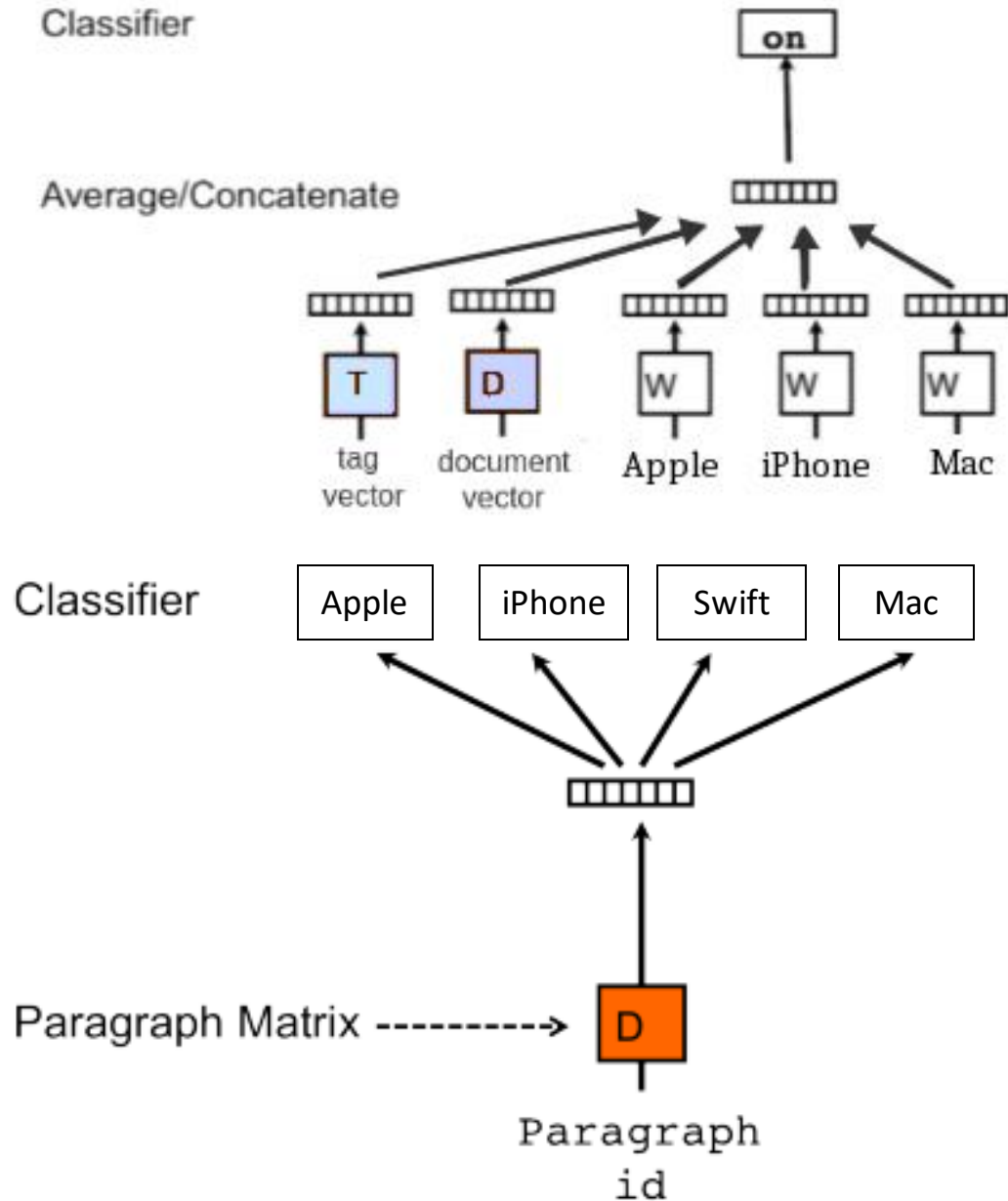
Trains a classifier
on the examples
it labeled

Use
Classifier

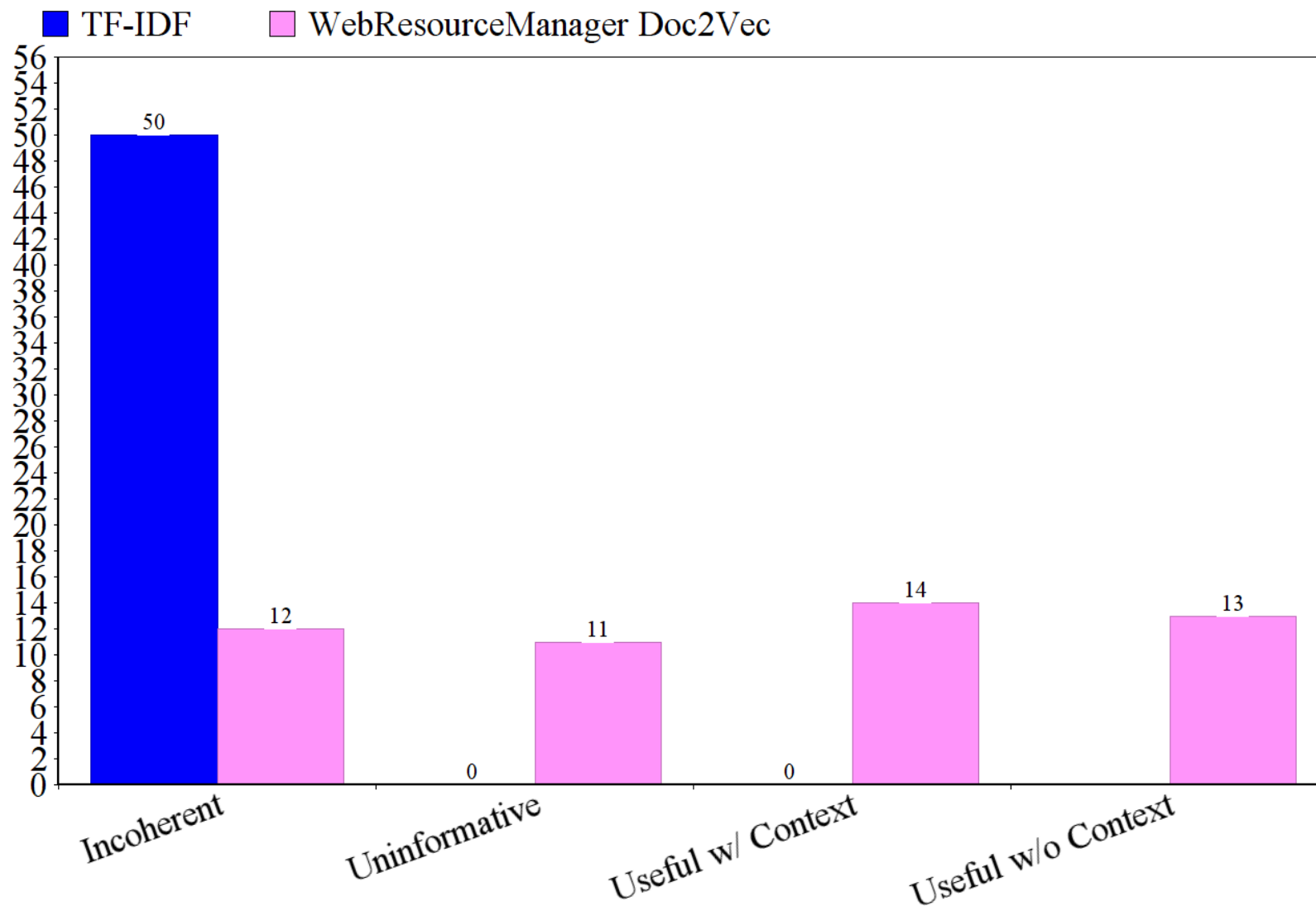
Classifies using
the features it
learned from
self-labeled data

Doc2Vec

- Represents semantic meaning of documents in a vector space
- You can "tag" documents with topics.
- We can attempt to cluster or classify documents using tags.



Classification Results: Web Pages



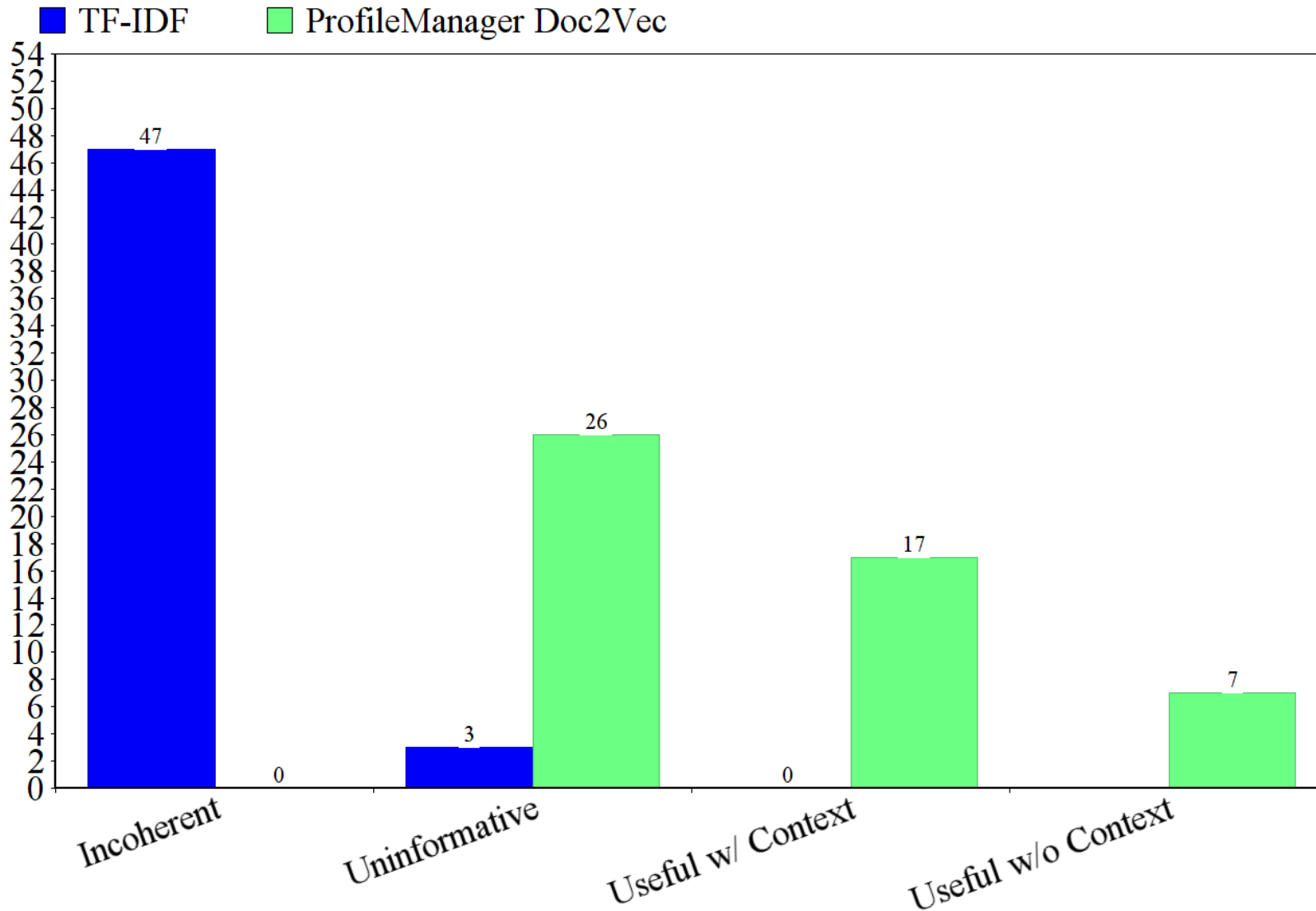
TF-IDF Produced:

- - riddel j
- 1941
- rhop
- danaida
- - boisduv j

We Produced:

- 2014 Chemr acquired 3D-Radar as a subsidiary of Curtiss-Wright Corporation in May 2014

Classification Results: Financial Statements

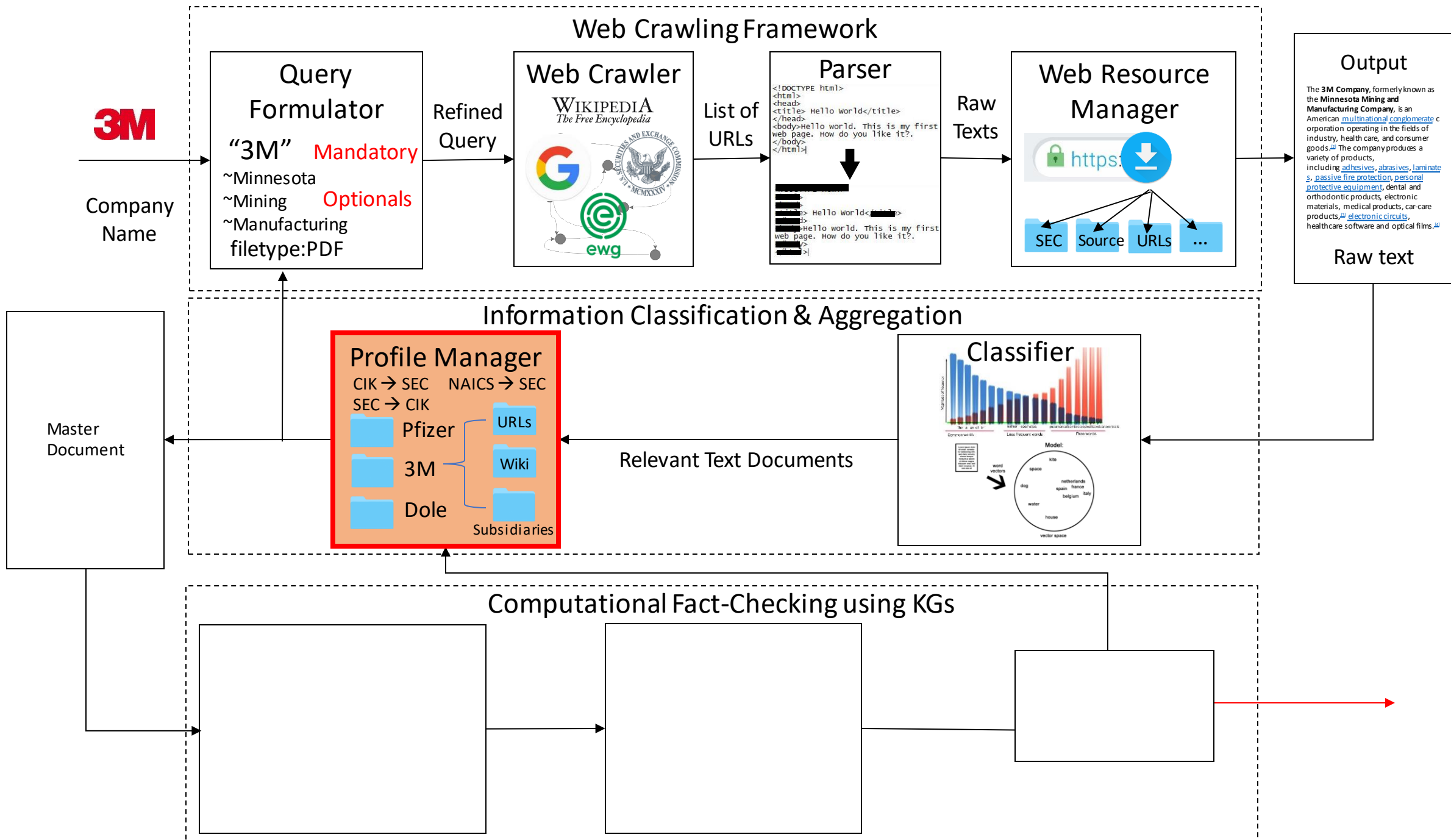


TF-IDF Produced:

- item 3
- asu no
- see note 2
- 10
- -11

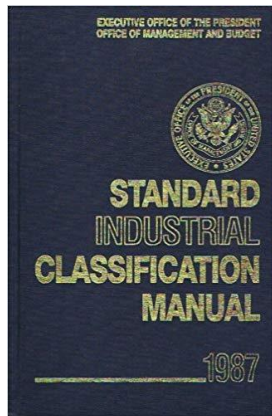
We Produced:

- these challenges add to the uncertainties of the legislative changes enacted as part of ACA





Central Index Key



Formerly 3M Company
3M wordmark.svg
Minnesota Mining and
Manufacturing Company (1902-
2002)

Type [Public](#)

Traded as

- [NYSE:MMM](#)
- [DJIA Component](#)
- [S&P 100 Component](#)
- [S&P 500 Component](#)

Industry [Conglomerate](#)
June 13, 1902; 116 years ago (as
Minnesota Mining and
Manufacturing Company)

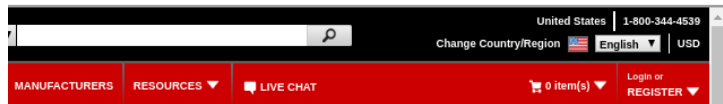
Founded [Two Harbors, Minnesota](#), U.S.

Founders John Dwan
Hermon Cable
Henry Bryan
William A. McGonagle

We found this list of subsidiaries:

[3M Scott Fire & Safety](#)
[Capital Safety](#)
[3M Japan Ltd](#)
[CUNO Inc](#)
[3M ESPE AG](#)
[3M Canada](#)
[3M India Ltd](#)
[3M Innovative Properties Company](#)
[3M Russia](#)
[3M United Kingdom PLC](#)
[3M A](#)
[3M Italia S.p.A.](#)
[Venture Tape Corp](#)
[3M Thailand Ltd](#)
[Aearo Technologies LLC](#)
[3M Nederland B.V.](#)
[3M Poland Sp z o.o.](#)
[3M Svenska AB](#)
[Scientific Anglers](#)
[3M Health Information Systems, Inc](#)
[Ceradyne](#)
[3M Touch Systems, Inc](#)
[3M Taiwan Ltd](#)
[3M China Ltd](#)
[3M Australia Pty. Ltd](#)
[3M Hong Kong Ltd](#)

<https://www.digikey.com/en/supplier-centers/3/3m>



These include 3M's Wiremount Insulation Displacement Contact (IDC), High Speed Hard Metric (HSHM) and the new Ultra Hard Metric (UHM)

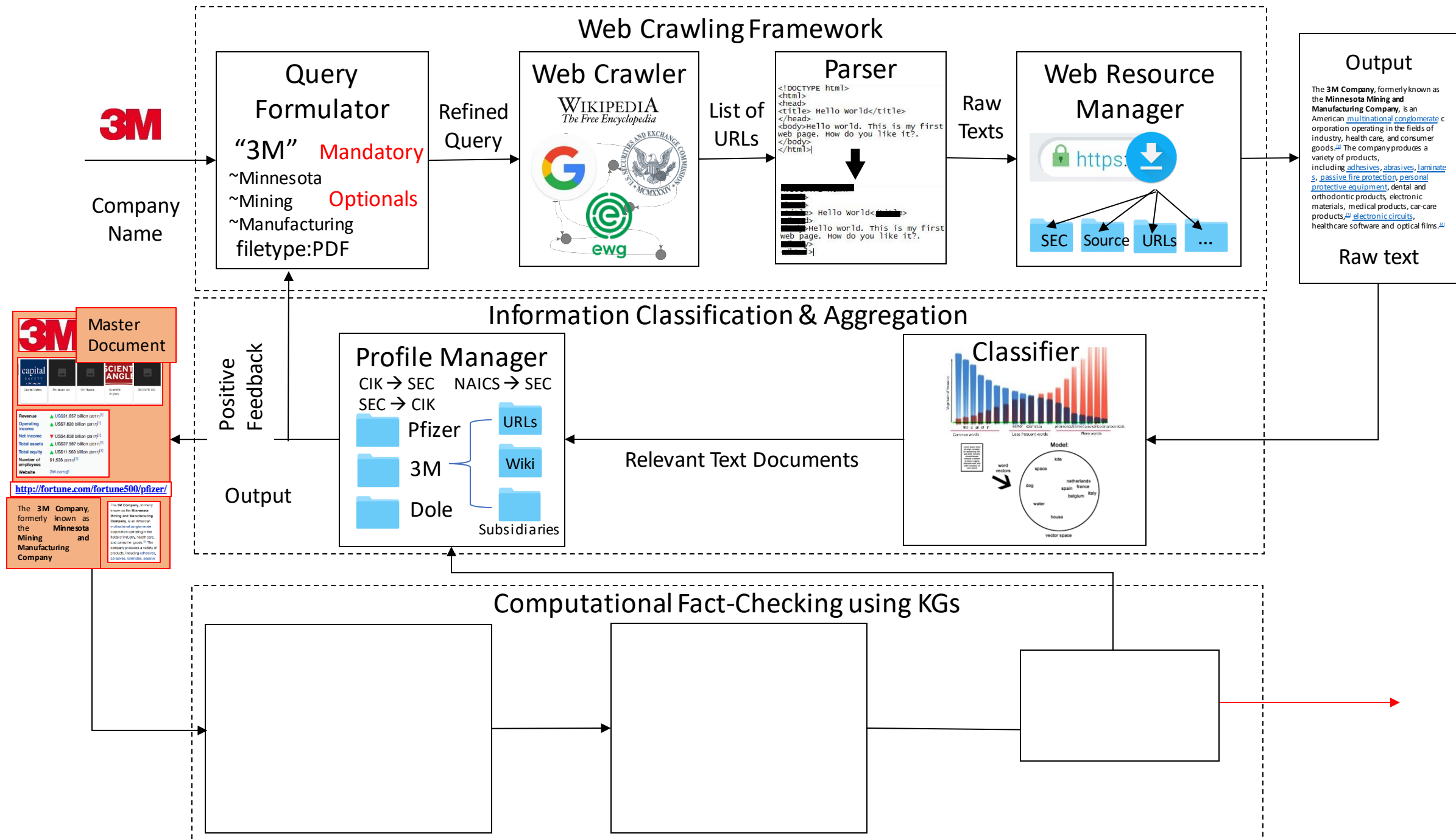
Using industry leading capabilities in CAD - such as NX and SolidWorks

+More 3M offers solutions for printed circuit board fabrication, board as sockets, carrier and cover tapes and trays, flexible circuits, and products

3M also offers solutions for shielding from EMI/RFI, for thermal management

Profile Manager

- Aggregates information by company
- Queryable
- Contains utility functions



Master Documents

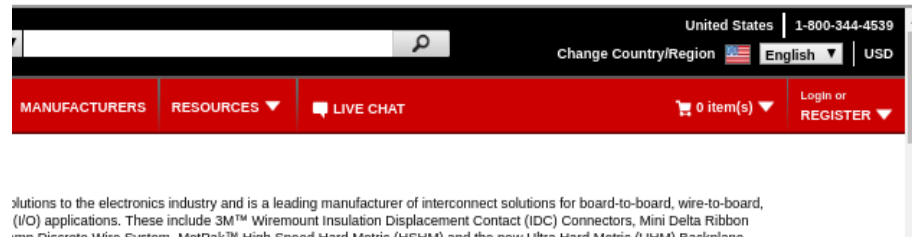
- Aggregates all the relevant company info
 - Wikipedia
 - Subidiaries
 - Web Crawler results
- Produced thousands for Praedicat and our code can produce as many as needed

	 3M Company Minnesota Mining and Manufacturing Company (1902-2002)
Formerly	
Type	Public
Traded as	<ul style="list-style-type: none">• NYSE:MMM• DJIA Component• S&P 100 Component• S&P 500 Component
Industry	Conglomerate
Founded	June 13, 1902; 116 years ago (as Minnesota Mining and Manufacturing Company) Two Harbors , Minnesota , U.S.
Founders	John Dwan Hermon Cable Henry Bryan William A. McGonagle

We found this list of subsidiaries:

[3M Scott Fire & Safety Capital Safety](#)
[3M Japan Ltd](#)
[CUNO Inc](#)
[3M ESPE AG](#)
[3M Canada](#)
[3M India Ltd](#)
[3M Innovative Properties Company](#)
[3M Russia](#)
[3M United Kingdom PLC](#)
[3M A](#)
[3M Italia S.p.A.](#)
[Venture Tape Corp](#)
[3M Thailand Ltd](#)
[Aearo Technologies LLC](#)
[3M Nederland B.V.](#)
[3M Poland Sp z o.o.](#)
[3M Svenska AB](#)
[Scientific Anglers](#)
[3M Health Information Systems, Inc](#)
[Ceradyne](#)
[3M Touch Systems, Inc](#)
[3M Taiwan Ltd](#)
[3M China Ltd](#)
[3M Australia Pty. Ltd](#)
[3M Hong Kong Ltd](#)

<https://www.digikey.com/en/supplier-centers/3/3m>

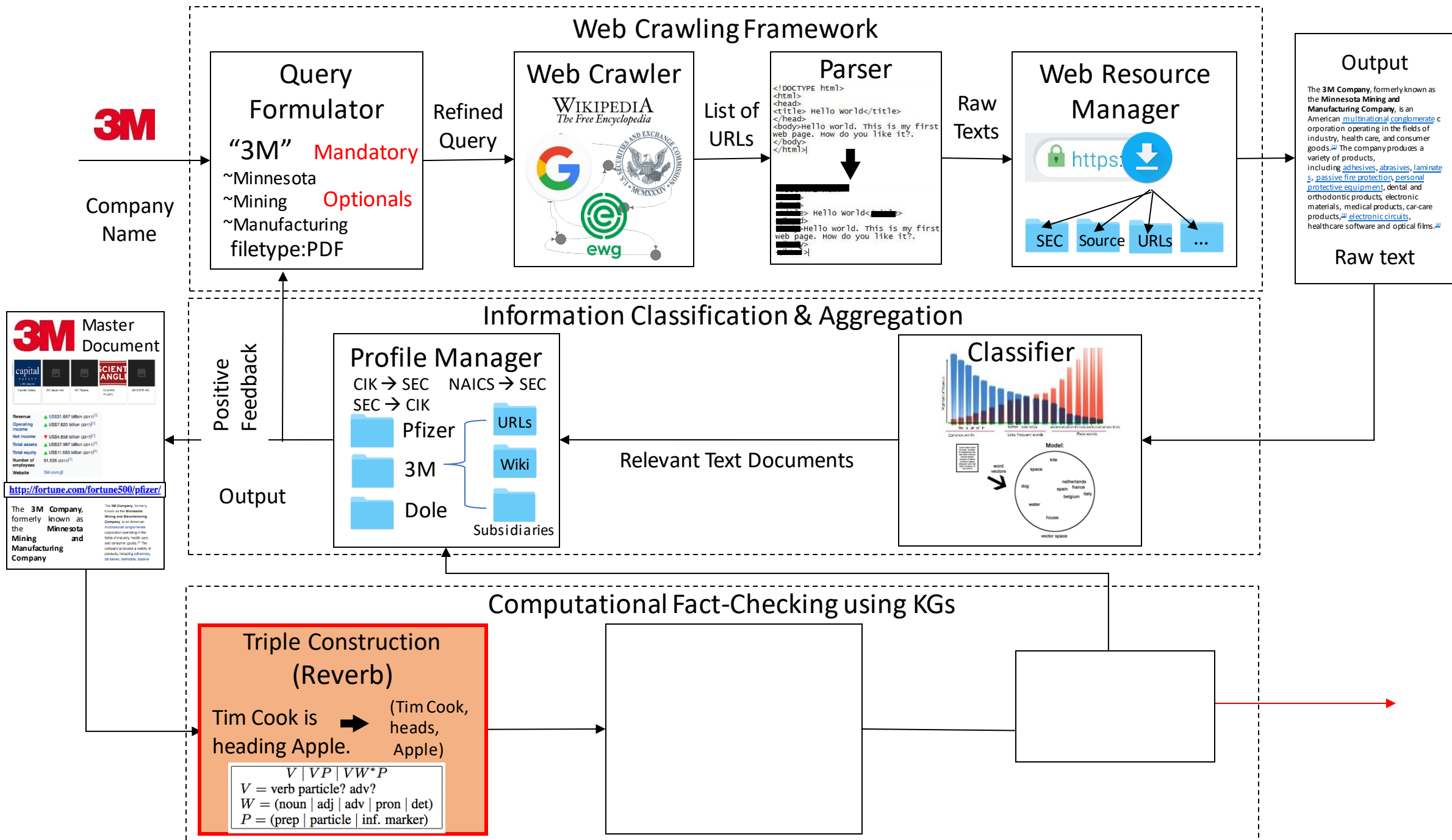


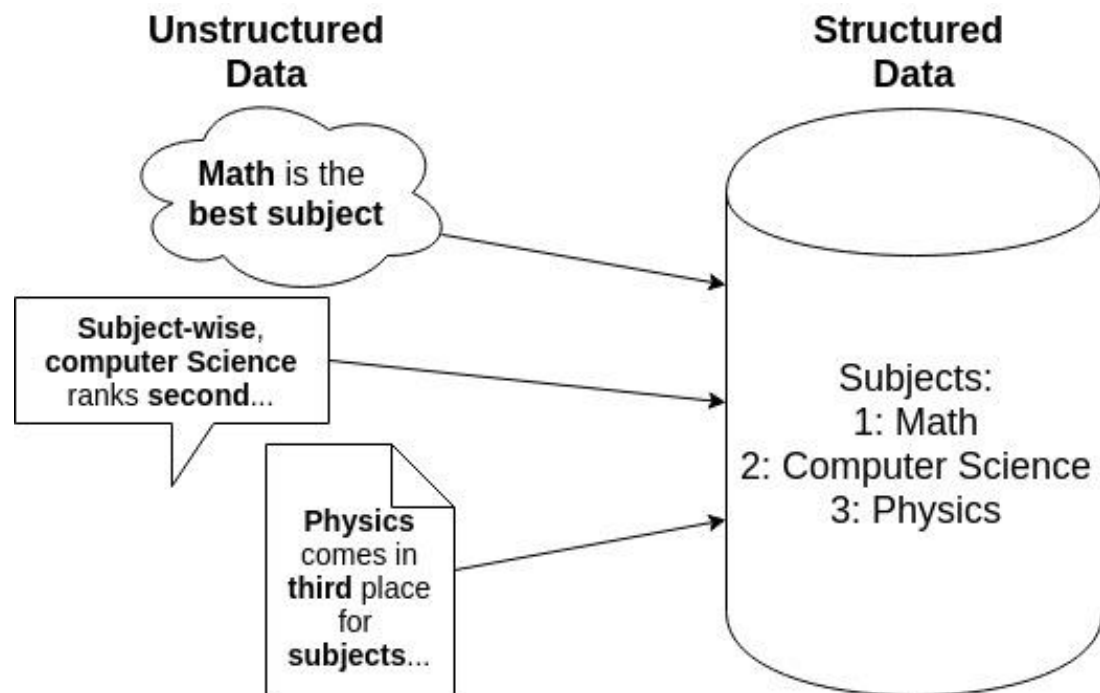
These include 3M's Wiremount Insulation Displacement Contact (IDC), High Speed Hard Metric (HSHM) and the new Ultra Hard Metric (UHM)

Using industry leading capabilities in CAD - such as NX and SolidWorks

+More 3M offers solutions for printed circuit board fabrication, board assembly, sockets, carrier and cover tapes and trays, flexible circuits, and products

3M also offers solutions for shielding from EMI/RFI, for thermal management





Open Information Extraction

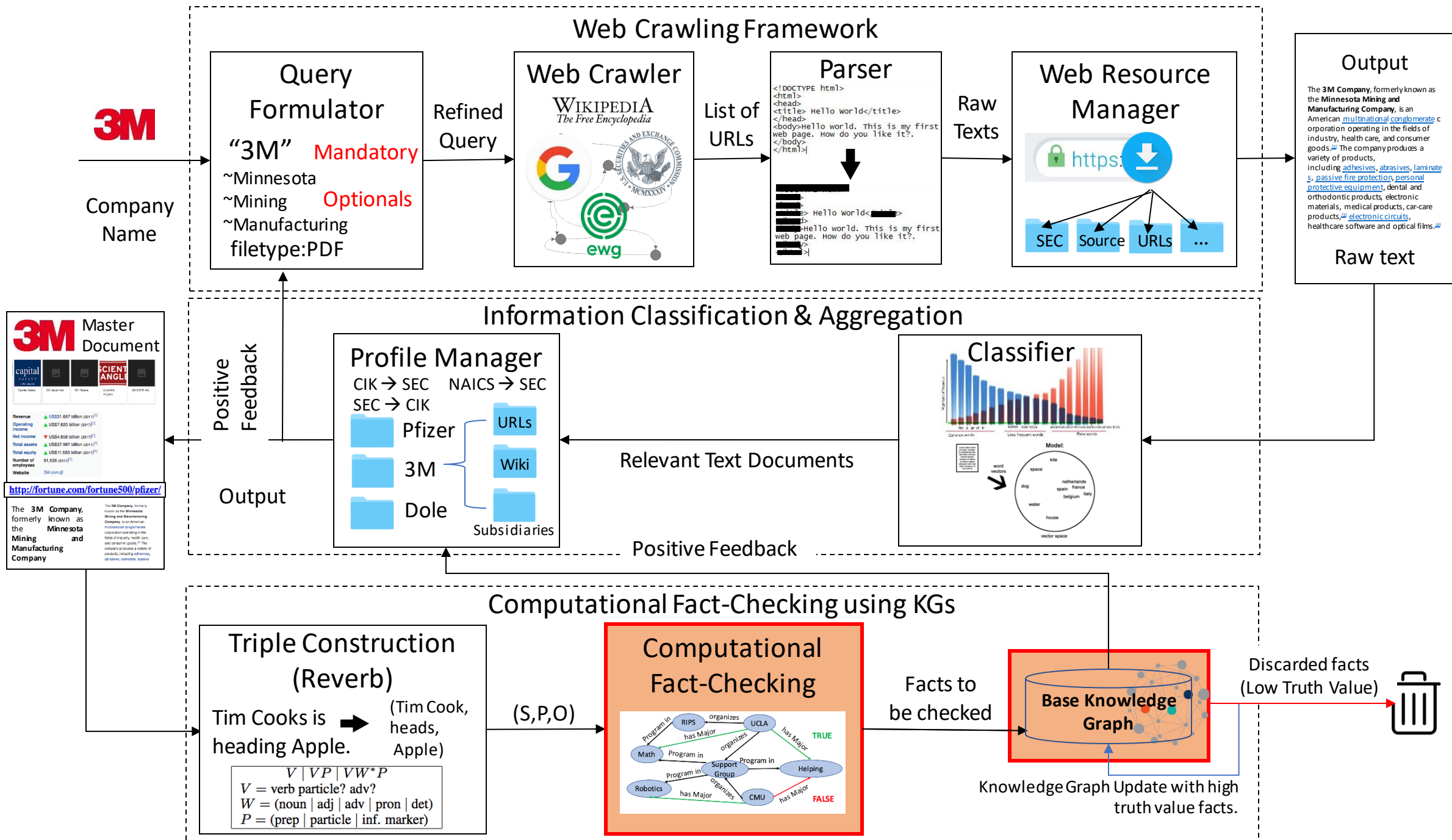
- Need to convert relevant text to structured data
- Reverb gives use this capability using Natural Language Processing

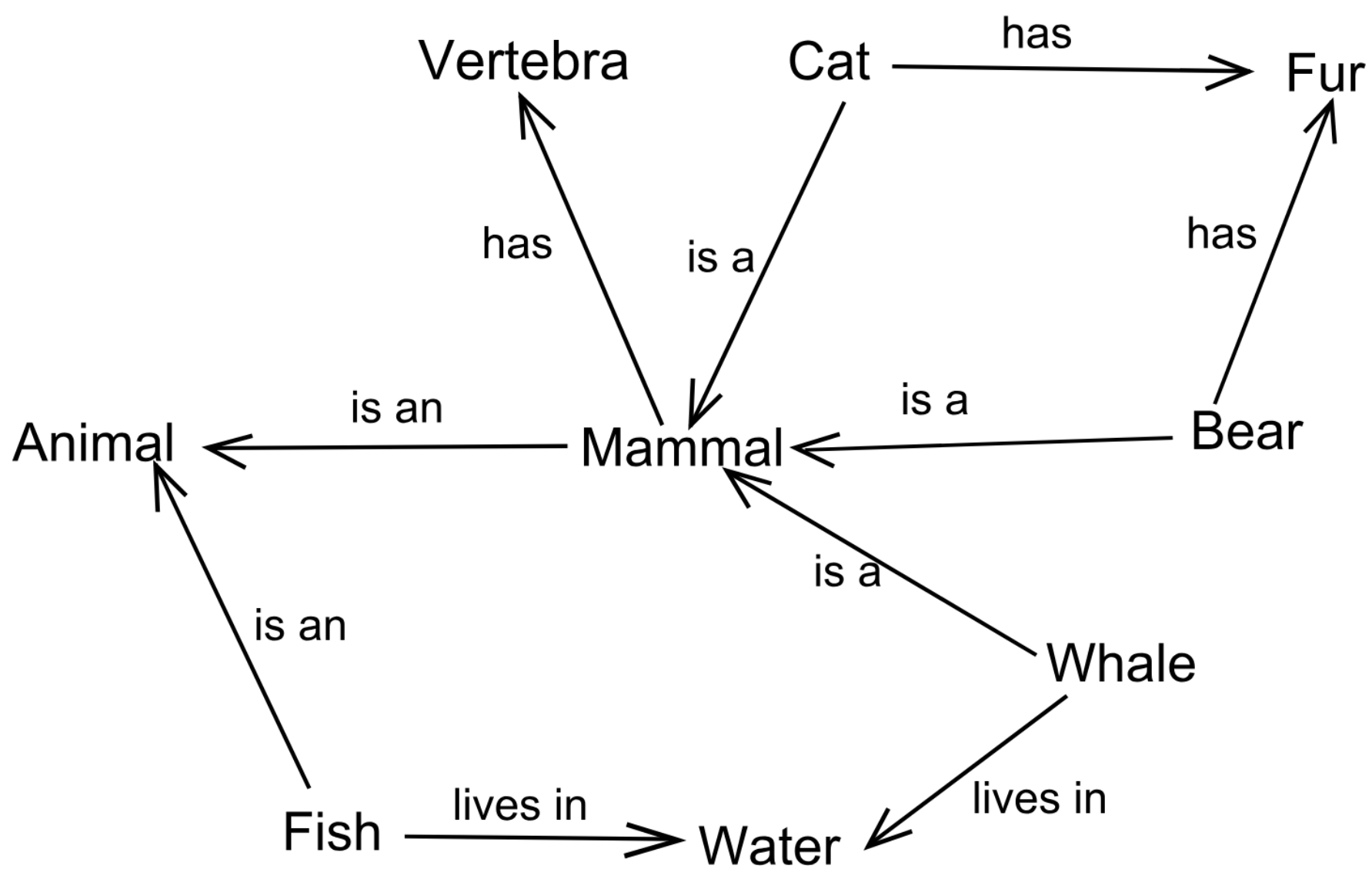
$V \mid VP \mid VW^*P$

V = verb particle? adv?

W = (noun | adj | adv | pron | det)

P = (prep | particle | inf. marker)





Knowledge Graphs

Knowledge Linker

- Valid facts should lie along specific paths

$$\mathcal{W}(\mathbf{P}_{s,o}) = \mathcal{W}(v_1, \dots, v_n) = \left[1 + \sum_{i=2}^{n-1} \log k(v_i) \right]^{-1}$$

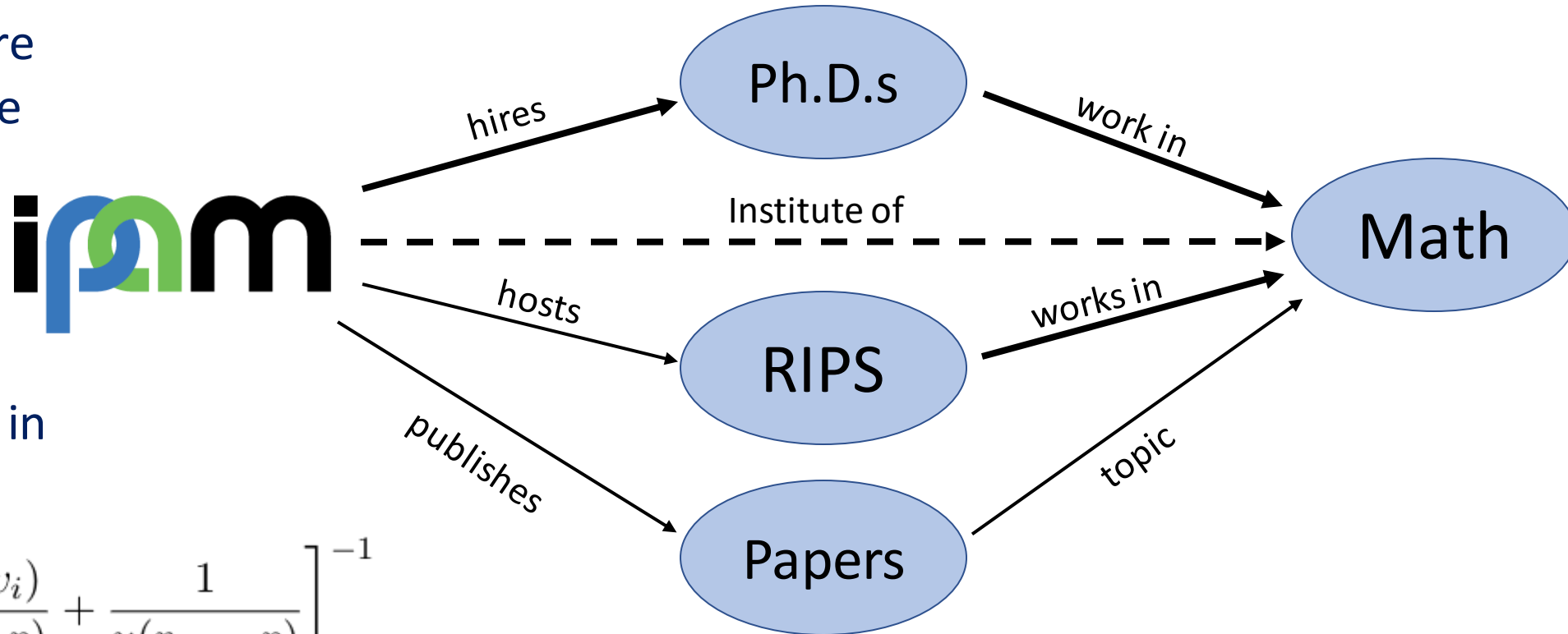


Knowledge Stream

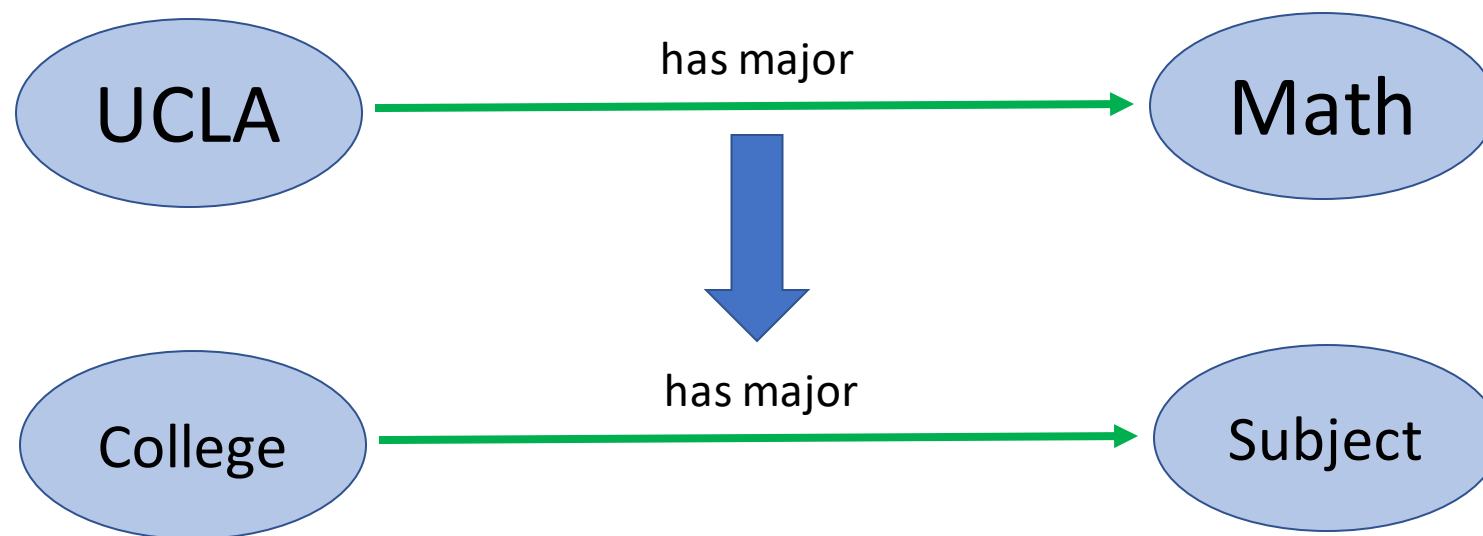
- A "stream" (set of paths) provides more context than a single path

- Relational similarity improves path specificity equation in Knowledge Linker

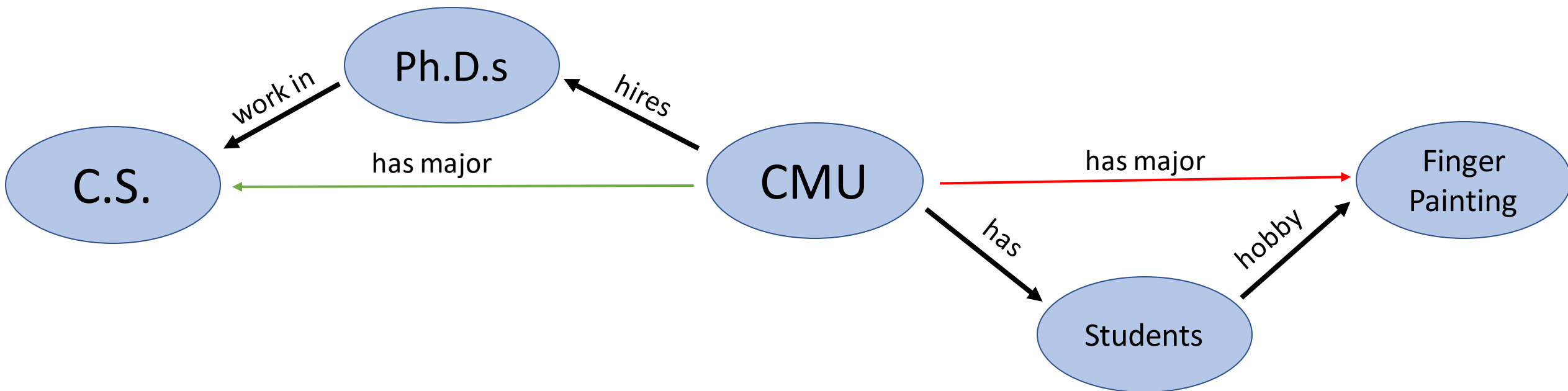
$$S'(P_{s,p,o}) = \left[\sum_{i=2}^{n-1} \frac{\log k(v_i)}{u(r_{i-1}, p)} + \frac{1}{u(r_{n-1}, p)} \right]^{-1}$$



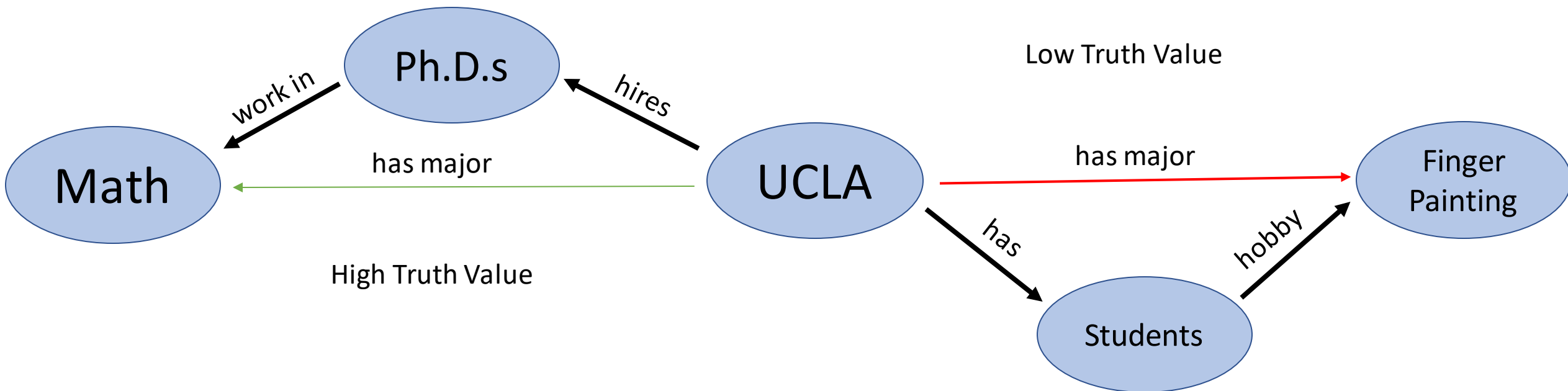
PredPath



PredPath

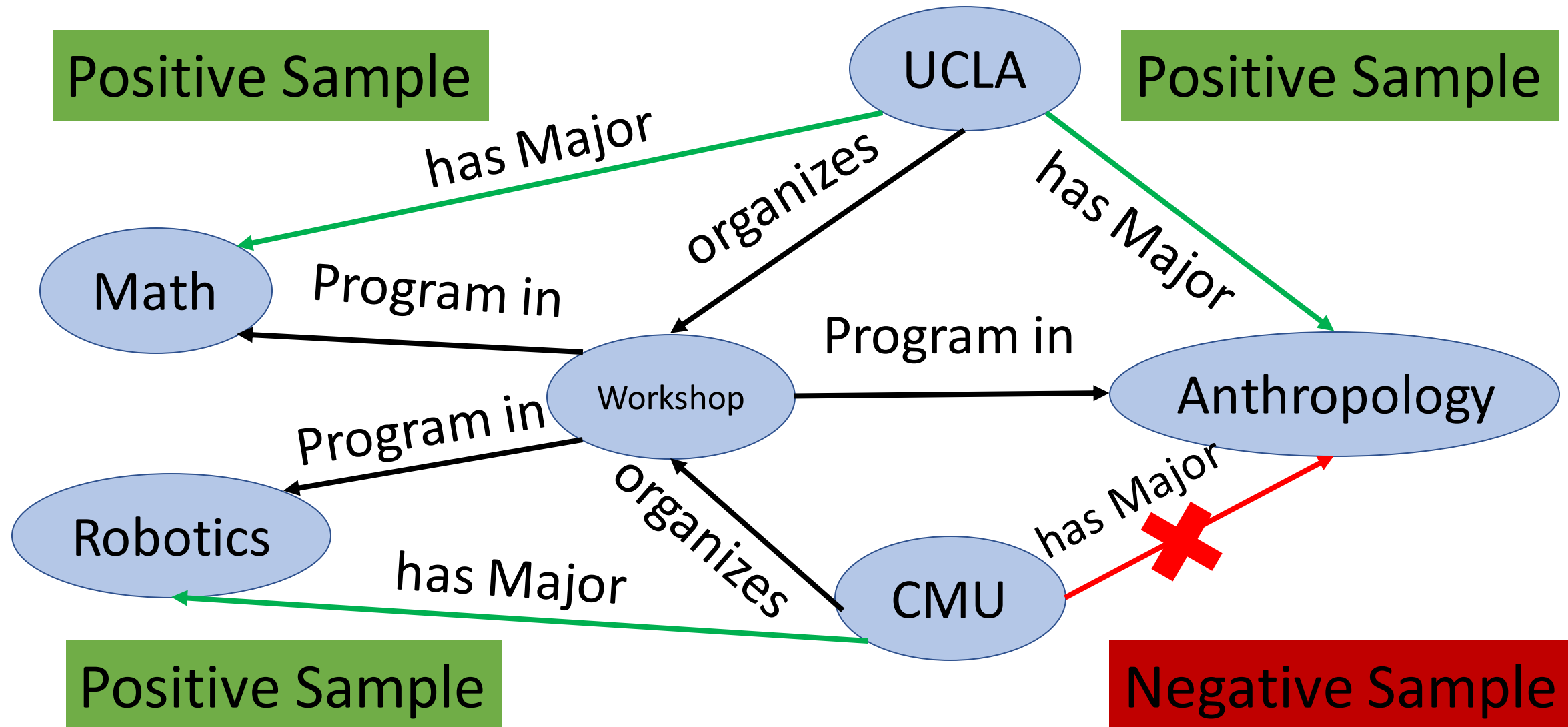


PredPath



Towards a New Computational Fact-Checking Algorithm

Why both negative and positive samples?



*StreamMiner, motivated by PredPath**

**Built negative and positive feature sets
for training on graphs.**

Path Specificity

$$\text{Path Specificity} = \text{Node Specificity} + \text{Path Similarity}$$


How general the idea of the node is
(how many concepts are connected to it)

Very General: University

Very Specific: Conference Room, IPAM, UCLA

How similar two relations are
e.g.: Mentors

Highly Similar: advises, counsels

Less similar: robs, steals

*StreamMiner, motivated by KREL-LINKER**

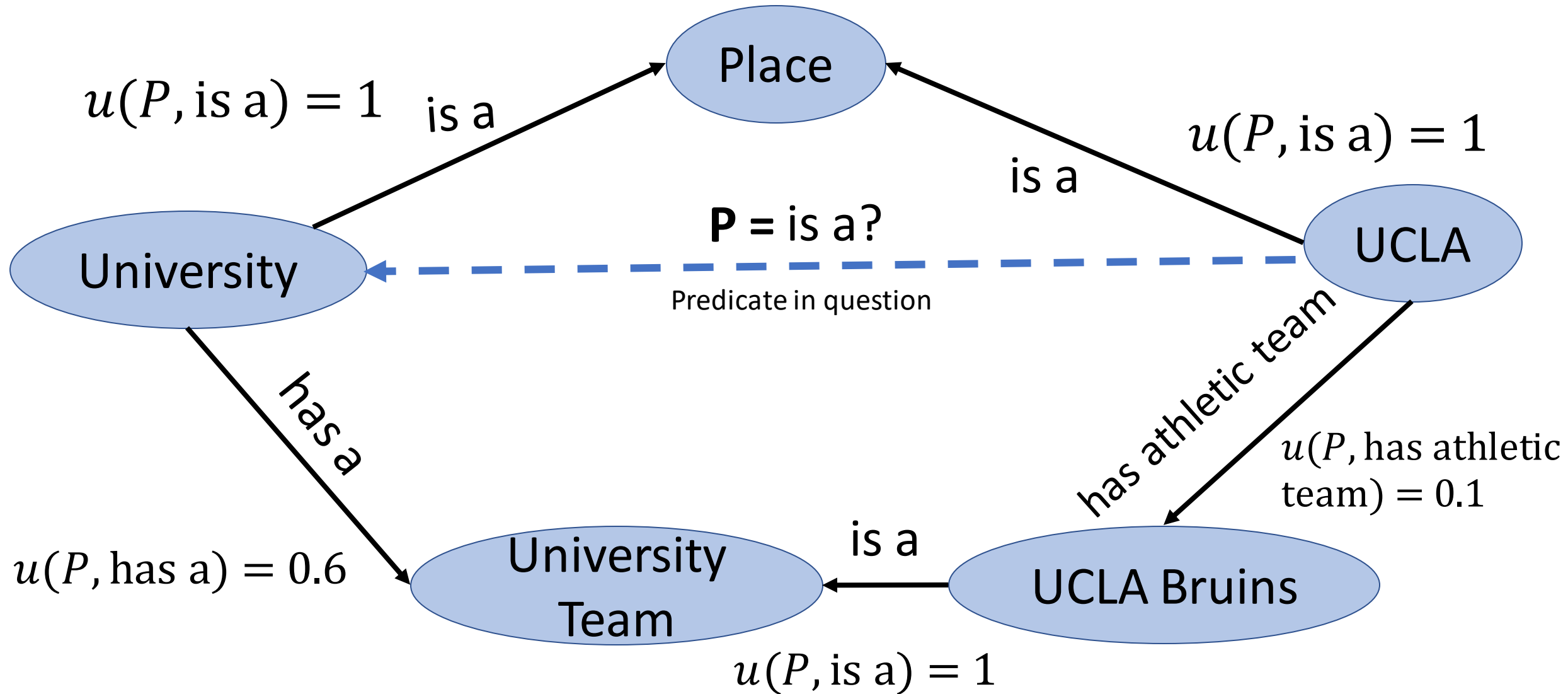
Path Specificity = Node Specificity + Path Similarity

Logarithm of
node in-degree

Relational similarity
w.r.t. predicate P as
cosine distance of co-
occurrence

$$S'(P_{s,p,o}) = \left[\sum_{i=2}^{n-1} \frac{\log k(v_i)}{u(r_{i-1}, p)} + \frac{1}{u(r_{n-1}, p)} \right]^{-1}$$

Path Specificity is more important than Path length

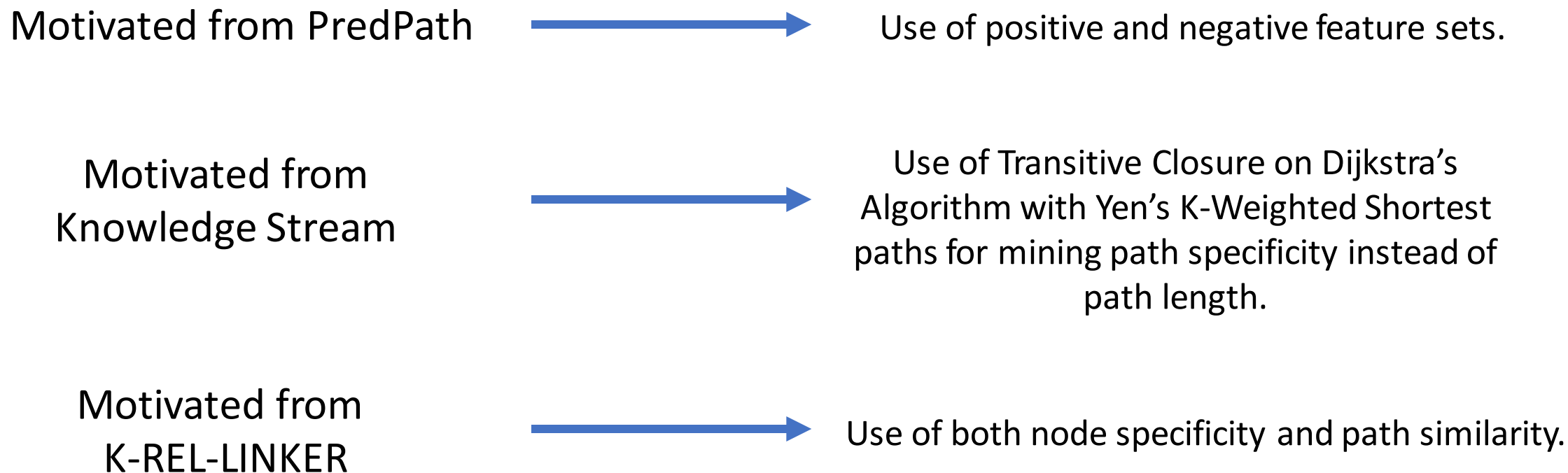


*StreamMiner, motivated by Knowledge Stream**

**Use of Transitive Closure on Dijkstra's Algorithm with
Yen's K-Shortest paths for mining path specificity
instead of path length.**

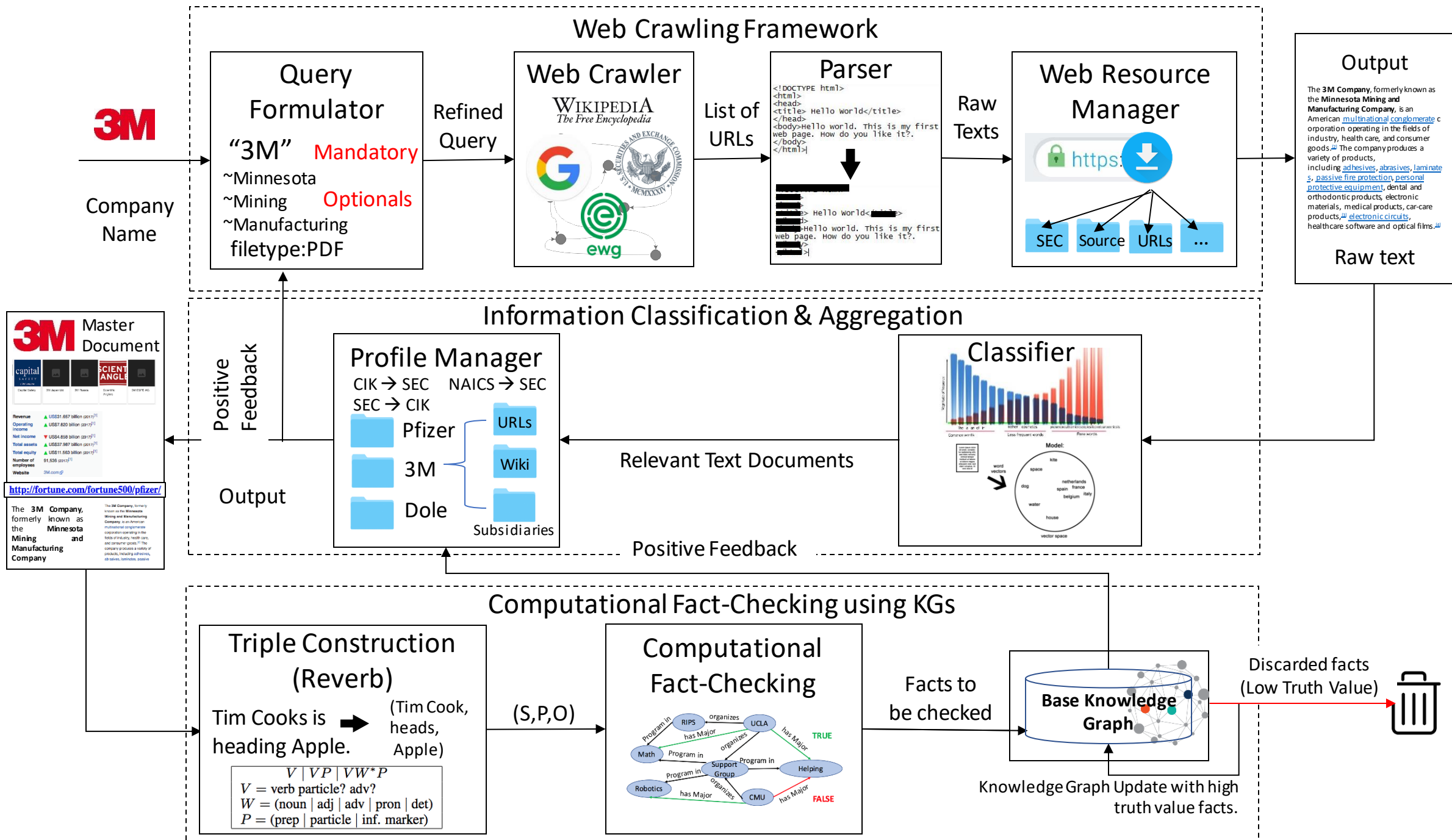
* P. Shiralkar, A. Flammini, F. Menczer, and G. L. Ciampaglia, *Finding streams in knowledge graphs to support fact checking*, CoRR, abs/1708.07239 (2017).

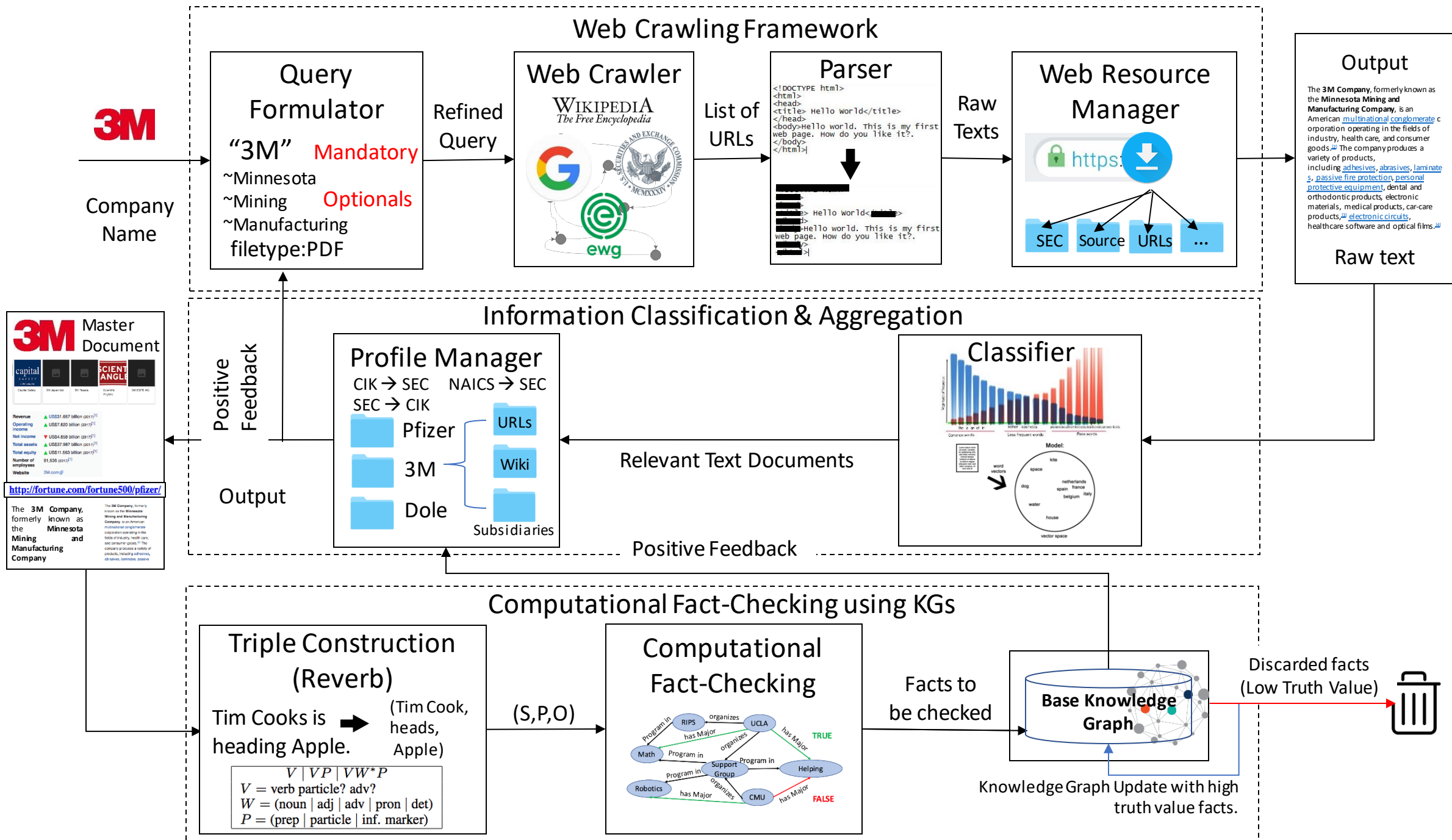
Stream Miner, Novel Fact Checking Algorithm



Stream Miner: Performance

Stream Miner was able to produce an average score of **86.325 (AUROC, Area under True Positive v/s False Positive Curve)** on a sub-sample database in its first run, which was at-par with the benchmark and state-of-the-art model PredPath.





Conclusion

Contributions

- A web crawling, classification and fact-checking architecture.
- A classification technique for retrieving relevant information.
- A fact-checking algorithm, StreamMiner, for checking information credibility.

Contribution: Making Impact



- Scaled up the Analysts' ability to retrieve information



- Data of 52,000+ Companies for decision-making

Acknowledgements



Shadi Shahsavari,
our academic mentor



Dr. Stephen DeSalvo,
Industry Mentor



Melissa Boudrea,
Industry Sponsor



Urjit Patel
Industry Mentor



Susana Serna,
our Program Director



David Medina,
Our IT Professional



Dimi Mavalski
Program Coordinator



Ronald McFarland
Program Co-ordinator



Questions?