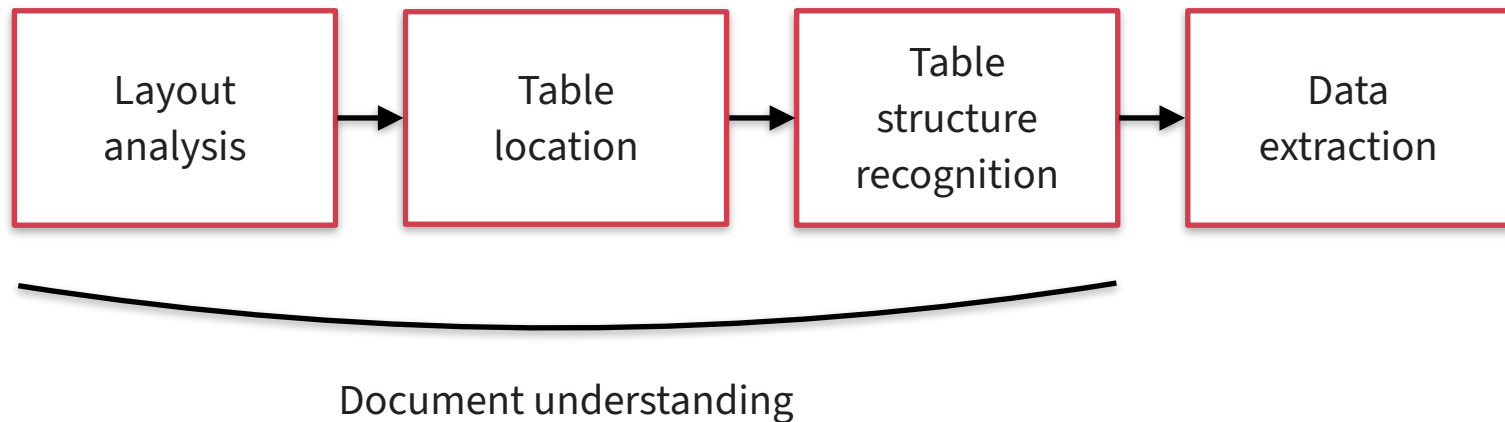


MISSOURI
1953

TOTAL PRECIPITATION AND DEPARTURES FROM NORMAL																										
Station	January		February		March		April		May		June		July		August		September		October		November		December		Annual	
	Precipitation	Departure	Precipitation	Departure	Precipitation	Departure	Precipitation	Departure	Precipitation	Departure	Precipitation	Departure	Precipitation	Departure	Precipitation	Departure	Precipitation	Departure	Precipitation	Departure	Precipitation	Departure	Precipitation	Departure	Precipitation	Departure
VANCE 5 ESE	2.71	0.98	1.47	2.34	6.08	1.12	3.98	1.19	3.01	1.26	2.12	2.18	1.46	1.90	.92	2.89	.21	2.43	2.21	0.36	1.14	2.12	1.59	1.31	26.89	17.84
LBANY	.16	0.67	2.01	0.39	3.38	1.69	3.07	0.78	2.84	0.16	2.40	2.64	.88	3.65	1.21	2.39	1.30	4.25	1.04	1.75	1.47	0.29	1.49	0.62	21.30	12.70
LTON 8 ENE	1.85		2.30		5.35		4.49		3.15		2.45		.43		.50		.15		1.05		2.06		1.43		25.23	
NITY	1.14	1.41	2.02	0.50	3.11	0.84	4.61	0.51	3.54	1.71	4.29	2.71	1.74	0.89	.55	3.65	.99	1.75	.54	1.57	1.61	1.07	2.55	0.51	25.73	12.40
NDERSON 1 SW	1.34	0.97	1.27	0.51	5.13	1.43	3.43	0.72	2.37	3.38	.47	5.46	2.24	2.17	1.76	2.36	1.23	2.65	3.04	0.85	.97	2.43	1.47	0.94	26.81	19.57
INDAPOLIS 7 SW	2.42		2.08		4.18		2.74		2.77		2.94		2.58		.36		.60		1.82		1.68		.83		25.09	
PPLETON CITY	1.37	0.47	.87	0.80	3.90	0.75	4.15	0.02	3.13	1.59	1.65	5.42	2.54	1.77	.78	3.18	2.93	1.45	2.81	0.15	.82	1.48	1.20	0.39	26.13	16.03
ACADEIA	2.16	0.83	1.30	1.18	4.52	0.47	3.60	2.22	2.14	2.10	3.04	1.10	1.88	2.24	.55	3.82	.30	3.68	1.84	1.61	1.38	1.74	.66	2.19	22.57	22.74
UYASSE	1.45	.95			4.05		2.83		3.60		3.28		2.73		1.60		2.35		2.78		.59		.91		27.13	
44 RANGER STA	1.68		.85		6.03		5.02		2.69		.19		1.19		.78		.33		2.95		1.06		1.38		24.15	
ELLEVIEV	1.24	1.06	1.02	1.81	4.23	0.42	2.75	2.26	2.29	2.82	3.16	1.44	.37	2.96	.92	1.85	.00	4.04	2.00	1.50	1.25	2.18	.61	2.20	19.83	23.70
ERNIE	2.14		2.69		6.85		3.80		4.13		1.98		4.75		.23		.91		1.73		1.26		2.76		33.23	
ETHANY	.22	0.97	2.15	0.73	3.60	1.37	3.87	0.58	2.52	1.55	5.32	0.26	1.59	2.57	1.15	2.67	.92	3.64	.15	2.77	1.16	0.88	1.44	0.29	26.09	11.92
IRCH TREE	2.08	1.10	1.88	0.82	5.72	1.74	4.39	0.37	3.44	0.91	2.00	2.41	.95	2.63	.95	3.81	.21	3.47	1.23	2.05	1.97	1.19	1.38	1.37	26.20	18.19
ACK 8 NW	2.14	0.56	1.46	0.82	3.70	0.19	4.24	0.65	2.53	1.98	4.74	0.68	2.10	1.59	1.29	2.95	.08	3.78	1.44	2.15	1.37	1.35	1.14	1.31	26.23	16.65

Data extraction pipeline



Structure of the presentation

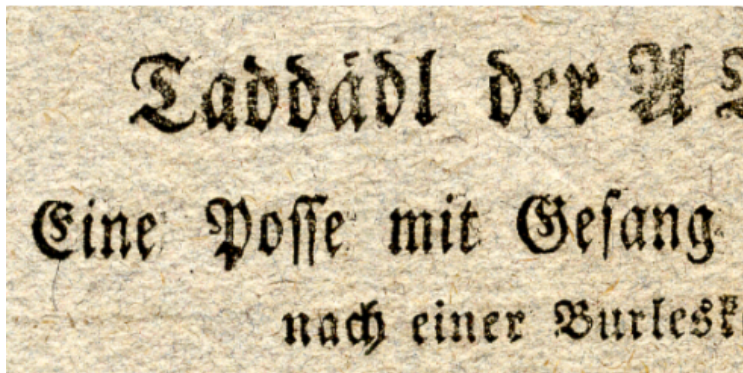
- Document understanding
 - Document image analysis
 - Object-level analysis of PDF
 - “Rule-based” systems
 - Machine learning
- Revisiting explainable algorithms
- Generative AI
- Comparison and conclusions

Document understanding



Document image analysis (1)

- Separation of foreground from background



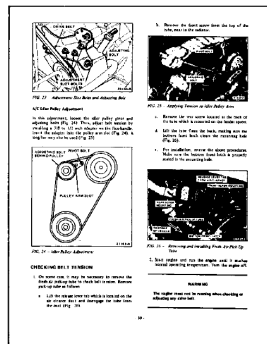
- Thresholding or binarization



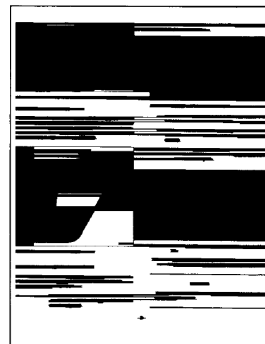
Source: Don Juan Archive, Vienna

Document image analysis (2)

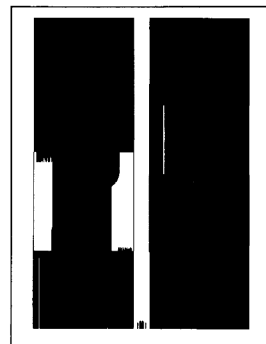
- Horizontal and vertical projection profiles
- Results in lines of text and blocks



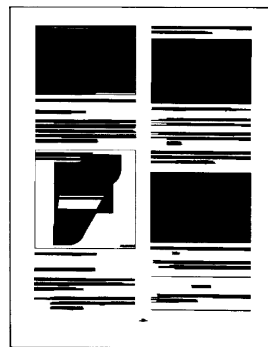
(a)



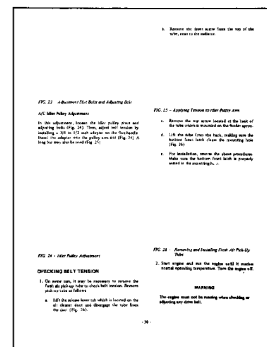
(b)



(c)



(d)

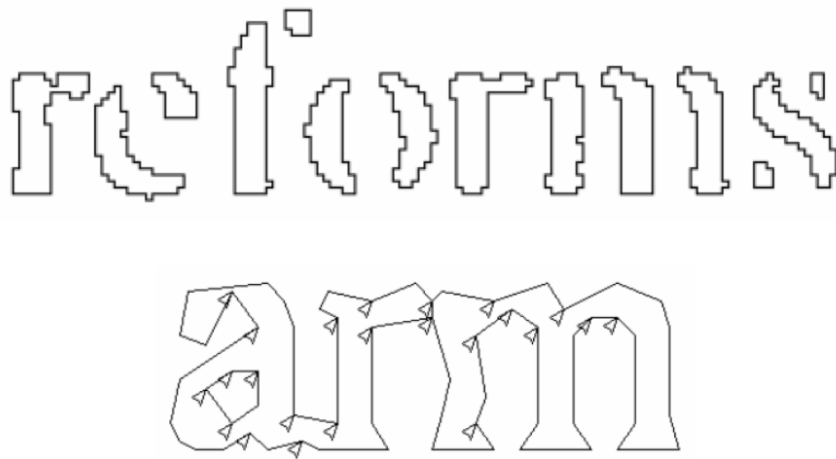


(e)

Source: Wahl et al.: Block Segmentation and Text Extraction in Mixed Text/Image Documents, CGIP 20 (4), pp. 375–390

Document image analysis (3)

- Segmentation into individual characters often poses problems



Source: Ray Smith: An Overview of the Tesseract Engine, ICAR 2007

Document image analysis (4)

- Multiple segmentation hypotheses
- Segmentation-free and window-based recognition methods



morgenländischen

Character classification

- A classifier, trained on ground-truthed data, returns the character code and corresponding confidence measure
- Typically convolutional neural networks are used
- Obtained words checked against the dictionary and/or linguistic rules
- Final result is the best combination of segmentation, classification and linguistic corrections

Processing PDF directly?

BT (begin text)

100 150 Td (move to coordinates)

[(T) 30 (his is a ty) 20 (pic) 40 (al sen) -10 (ten) -10 (ce.)] TJ

ET (end text)

[(AWAY again)] TJ	AWAY again
[(A) 120 (W) 120 (A) 95 (Y again)] TJ	AWAY again

Text fragments example

Wer etwas ist oder sein möchte bei Daimler-Benz, der achtet auf die Kleiderordnung; man trägt Blau im Schwabenkonzern, hell am Fließband, dunkel auf der Führungsebene. Und wer den Entscheidungsträgern im Vorstand ganz nahe ist, der darf sich OFK-Mitglied nennen, der gehört zum oberen Führungskreis des Hauses.

Ende Januar zogen rund 1000 der 1400 OFK-Mitglieder in die Stuttgar

Object-based layout analysis

UPDATE

In other news

- In former Soviet states, leaders watch uneasily as the call for democracy widens. *Page 3*
- Rebellion at La Scala ends as Riccardo Muti resigns as musical director after 19 years. *Page 8*
- A UN envoy says Syria has vowed to pull out all military and intelligence units from Lebanon by the end of April. *Page 9*
- Iraqi lawmakers elect a Sunni Arab as speaker of Parliament. *Page 9*

Daring to create

Mayor Bertrand Delanoë has worked hard to make his city a contemporary masterpiece. "Paris is a museum, and that is a privilege," he says. "But if it wants to be loyal to its history, it needs to innovate, to dare — it needs to move into the 21st century." *Page 2*



On the Web: www.iht.com

Das Windsor-Syndrom

„Fordern Sie uns! Grillen Sie uns!“, appellierte Konzernchef Jürgen Schrepp an seine Führungsräfte. Auf dem Topmanagement-Meeting konnte er auch die Attacken seines Vorgängers Edzard Reuter

Markt, und seine Division soll 1998 zum erstenmal wieder schwarze Zahlen schreiben. Als Nachfolger wird der neue Mann des Bereichs Lkw-Antriebsstrang, **Klaus Müller** (44), gebührt.

Durch rechtzeitige Pensionierung entzieht sich **Peter Fietzek** (59), Bereichsvorstand und Asien-Beauftragter, den Folgen der Reorganisation in Frankfurt.

Künftig regieren in der Region vier „Chief Executives“ (Japan, Indochina, Asien, China), die jeweils das gesamte Konzerngeschäft von

Wer etwas ist oder sein möchte bei Daimler-Benz, der achtet auf die Kleiderordnung: man trägt Blau im Schwebenkonzept, hell am Fließband, dunkel auf der Führungsebene. Und wer den Entscheidungsträgern im Vorstand ganz nahe ist, der darf sich OFK-Mitglied nennen, der gehört zum oberen Führungskreis des Hauses.

Ende Januar zogen rund 1000 der 1400 OFK-Mitglieder in die Stuttgarter Liederhalle ein, viele dunkelblau gekleidet und alle gespannt wie Chöre auf einen großen Auftritt: **Jürgen Schrepp** (53), der Chef, hatte gefunzt, und er verlangte Manuskript: „Dies ist unsere gemeinsame Veranstaltung. Nennen Sie sie! Fordern Sie uns! Grillen Sie uns!“

Für Spannung bei dem Treffen am 27. Januar war gesorgt: Nie zuvor lag ein Mecklerfolg und Meinungsgegnert so nahe beieinander, nie zuvor rankten sich so viele Gerüchte um Vorfälle; nie zuvor hatte ein ehemaliger Vorsitzender so mit dem Unternehmen abgerechnet wie jetzt **Edzard Reuter** (70) in seinem Monolog (siehe Kasten Seite 16).

Was sagt Schrepp intern im persönlichen Brief-Panorama, was zum vergrößerten Senat? Welche Signale sendet der Vorstandschef in Richtung seiner Kollegen **Jürgen Hubbert** (58, Pkw-Geschäft) und **Dieter Zetsche** (44, früher Entwicklung, heute Vertrieb), die für das Debakel die Verantwortung tragen?

Der Menschheitskognitivist ließ sich nicht aus der Reserve locken: „Wir sitzen Ihnen klar als Team gegenüber“, verkündete Schrepp. Ein großes Reversenwort, so es denn dann kommt, blüht bis nach der Hauptversammlung am 27. Mai raus.

Dennoch gab es am Rande des Gipfelfestivals zwei Toppenationen: Der künftige **Horst Zimmer** (62), der es als omnipotenter Chef des Geschäftsbereichs Lkw-Europa immer



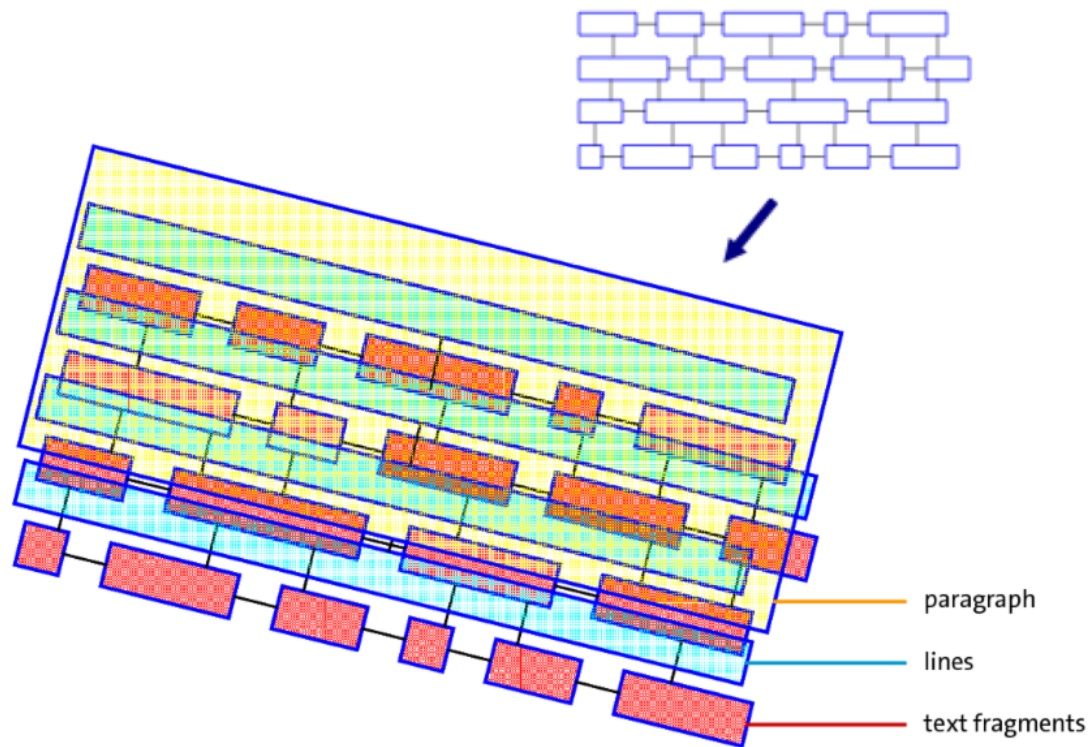
ziemlich egal war, wer unter ihm den Nachfolgeposten abgab, schaltete am Konzern.

Zimmer geht aus gesundheitlichen Gründen ins Zenit seiner Karriere: Die beiden neuen Modellfreibahn Actros (schwere Lastkraftwagen) und Actros (Chassis-Lkw) sind im

Flugmodus bis Finanzdienstleistungen verantwortlich.

Auf der Veranstaltung selbst ging es dann zunächst um einen Mann, der seit knapp drei Jahren im interner Daimler-Benz nicht mehr eingefallen wird: **Edzard Reuter**. Das Wirken seines Vorgängers, kann Schrepp ab

Graph-based multi-level representation



Potential problems

- Text does not have to be in reading order
- Operators intended for kerning can be used to jump across columns
- Character encoding may be unknown
- Overprinting to simulate boldness
- and, of course, text may be part of an image object (or not recognized by OCR...)

Matrox product selection table

	Axio LE	Axio HD	Axio SD	RTX2
Capture/editing formats				
HDV 1080i	X	X	X	X
HDV 720p	X	X	X	next release
DVCPRO HD	X	X	X	—
DV, DVCPRO, DVCAM	X	X	X	X
DVCPRO50	X	X	X	—
P2 MXF – DVCPRO50, DVCPRO HD	X	X	X	—
XDCAM MXF – DVCAM, IMX	X	X	X	—
XDCAM HD MXF 18, 25, 35 mbps	X	X	X	—
Slow & Quick Motion				
MPEG-2 4:2:2 I-frame SD*	10-50 mbps	10-50 mbps	10-50 mbps	10-25 mbps
Uncompressed 8-bit SD*	X	X	X	—

Table recognition (1/4)

Phosphorus (yellow or white)	7723-14-0	<u>Only</u> if it is a yellow or white form.
Sulfuric acid (acid aerosols including mists, vapors, gas, fog, and other airborne forms of any particle size)	7664-93-9	<u>Only</u> if it is an aerosol form as defined.
Vanadium (except when contained in an alloy)	7440-62-2	<u>Except</u> if it is contained in an alloy.
Zinc (fume or dust)	7440-66-6	<u>Only</u> if it is in a fume or dust form.

The qualifier for the following three chemicals is based on the chemical activity rather than the form of the chemical. These chemicals are subject to EPCRA section 313 reporting requirements only when the indicated activity is performed.

<u>Chemical/ Chemical Category</u>	<u>CAS Number</u>	<u>Qualifier</u>
Dioxin and dioxin-like compounds (manufacturing; and the processing or otherwise use of dioxin and dioxin-like compounds if the dioxin and dioxin-like compounds are present as contaminants in a chemical and if they were created during the manufacture of that chemical.)	NA	<u>Only</u> if they are manufactured at the facility; or are processed or otherwise used when present as contaminants in a chemical but only if they were created during the manufacture of that chemical.
Isopropyl alcohol (only persons who manufacture by the strong acid process are subject, no supplier notification)	67-63-0	<u>Only</u> if it is being manufactured by the strong acid process. Facilities that process or otherwise use isopropyl alcohol are <u>not</u> covered.
Saccharin (only persons who manufacture are subject, no supplier notification)	81-07-2	<u>Only</u> if it is being manufactured.

There are no supplier notification requirements for isopropyl alcohol and saccharin since the processors and users of these chemicals are not required to report. Manufacturers of these chemicals do not need to notify their customers that these are reportable EPCRA

After candidate column finding (2/4)

Phosphorus (yellow or white)	7723-14-0	<u>Only</u> if it is a yellow or white form.
Sulfuric acid (acid aerosols including mists, vapors, gas, fog, and other airborne forms of any particle size)	7664-93-9	<u>Only</u> if it is an aerosol form as defined.
Vanadium (except when contained in an alloy)	7440-62-2	<u>Except</u> if it is contained in an alloy.
Zinc (fume or dust)	7440-66-6	<u>Only</u> if it is in a fume or dust form.

The qualifier for the following three chemicals is based on the chemical activity rather than the form of the chemical. These chemicals are subject to EPCRA section 313 reporting requirements only when the indicated activity is performed.

<u>Chemical/ Chemical Category</u>	<u>CAS Number</u>	<u>Qualifier</u>
Dioxin and dioxin-like compounds (manufacturing; and the processing or otherwise use of dioxin and dioxin-like compounds if the dioxin and dioxin-like compounds are present as contaminants in a chemical and if they were created during the manufacture of that chemical.)	NA	<u>Only</u> if they are manufactured at the facility; or are processed or otherwise used when present as contaminants in a chemical but only if they were created during the manufacture of that chemical.
Isopropyl alcohol (only persons who manufacture by the strong acid process are subject, no supplier notification)	67-63-0	<u>Only</u> if it is being manufactured by the strong acid process. Facilities that process or otherwise use isopropyl alcohol are <u>not</u> covered.
Saccharin (only persons who manufacture are subject, no supplier notification)	81-07-2	<u>Only</u> if it is being manufactured.

There are no supplier notification requirements for isopropyl alcohol and saccharin since the processors and users of these chemicals are not required to report. Manufacturers of these chemicals do not need to notify their customers that these are reportable EPCRA

After horizontal merging (3/4)

Phosphorus (yellow or white)	7723-14-0	<u>Only</u> if it is a yellow or white form.
Sulfuric acid (acid aerosols including mists, vapors, gas, fog, and other airborne forms of any particle size)	7664-93-9	<u>Only</u> if it is an aerosol form as defined.
Vanadium (except when contained in an alloy)	7440-62-2	<u>Except</u> if it is contained in an alloy.
Zinc (fume or dust)	7440-66-6	<u>Only</u> if it is in a fume or dust form.

The qualifier for the following three chemicals is based on the chemical activity rather than the form of the chemical. These chemicals are subject to EPCRA section 313 reporting requirements only when the indicated activity is performed.

<u>Chemical/ Chemical Category</u>	<u>CAS Number</u>	<u>Qualifier</u>
Dioxin and dioxin-like compounds (manufacturing; and the processing or otherwise use of dioxin and dioxin-like compounds if the dioxin and dioxin-like compounds are present as contaminants in a chemical and if they were created during the manufacture of that chemical.)	NA	<u>Only</u> if they are manufactured at the facility; or are processed or otherwise used when present as contaminants in a chemical but only if they were created during the manufacture of that chemical.
Isopropyl alcohol (only persons who manufacture by the strong acid process are subject, no supplier notification)	67-63-0	<u>Only</u> if it is being manufactured by the strong acid process. Facilities that process or otherwise use isopropyl alcohol are <u>not</u> covered.
Saccharin (only persons who manufacture are subject, no supplier notification)	81-07-2	<u>Only</u> if it is being manufactured.

There are no supplier notification requirements for isopropyl alcohol and saccharin since the processors and users of these chemicals are not required to report. Manufacturers of these chemicals do not need to notify their customers that these are reportable EPCRA

Final result (4/4)

airborne forms of any particle size)		as defined.
Phosphorous (yellow or white)	7723-14-0	Only if it is a yellow or white form.
Sulfuric acid (acid aerosols including mists, vapors, gas, fog, and other airborne forms of any particle size)	7664-93-9	Only if it is an aerosol form as defined.
Vanadium (except when contained in an alloy)	7440-62-2	Except if it is contained in an alloy.
Zinc (fume or dust)	7440-66-6	Only if it is in a fume or dust form.

The qualifier for the following three chemicals is based on chemical activity rather than the form of the chemical. These chemicals are subject to EPCRA section 313 reporting requirements only when the indicated activity is performed.

Chemical/ Chemical Category	CAS Number	Qualifier
Dioxin and dioxin-like compounds (manufacturing; and the processing or otherwise use of dioxin and dioxin-like compounds if the dioxin and dioxin-like compounds are present as contaminants in a chemical and if they were created during the manufacture of that chemical.)	NA	Only if they are manufactured at the facility; or are processed or otherwise used when present as contaminants in a chemical but only if they were created during the manufacture of that chemical.
Isopropyl alcohol (only persons who manufacture by the strong acid process are subject, no supplier notification)	67-63-0	Only if it is being manufactured by the strong acid process. Facilities that process or otherwise use isopropyl alcohol are not covered.
Saccharin (only persons who manufacture are subject, no supplier notification)	81-07-2	Only if it is being manufactured.

There are no supplier notification requirements for isopropyl alcohol and saccharin since the processors and users of these chemicals are not required to report. Manufacturers of these chemicals do not need to notify their customers that these are

Limitations of conventional approaches

- “Knowledge gap:” Rules are inflexible and work in isolation on one level of granularity
- Difficult to backtrack

off Case	G. F Holder III (Hood III)*1	G.F Holder IV (Hood V) *1
_P1319	NC	(0)
_P1319	NC	(0)
_P1319	NC	(0)
_P1214	NC	(0)
_P1116	NC	(0)
IP814	(0)	(0)

isoners is
ny and the
orted.

Congress
ly six pris-
e, but that
hat the re-
al Albert
losed, sub-
s of last

6 prisoner
or suspec-
provided by
k after re-
reviewed
vestigators
iries and
criminal
navy offi-
still under

leaving court in New York on Tuesday
after he was convicted of all charges.

technology
The verc

CURRENCIES | New York

	Tuesday 4 P.M.	Previous
€1 =	\$1.3315	\$1.3369
£1 =	\$1.9128	\$1.914
\$1 =	¥104.49	¥104.925
\$1 =	SF1.1645	SF1.1594

Full currency rates | Page 14

OIL | New York

	Tuesday 4 P.M.
Light sweet crude	\$55.05  \$0.10

Pus

EU wa

By James F

PARIS: T
hind the cl
London's v
that who
Deutsche F
States wor
biggest ma

Machine learning for structure detection

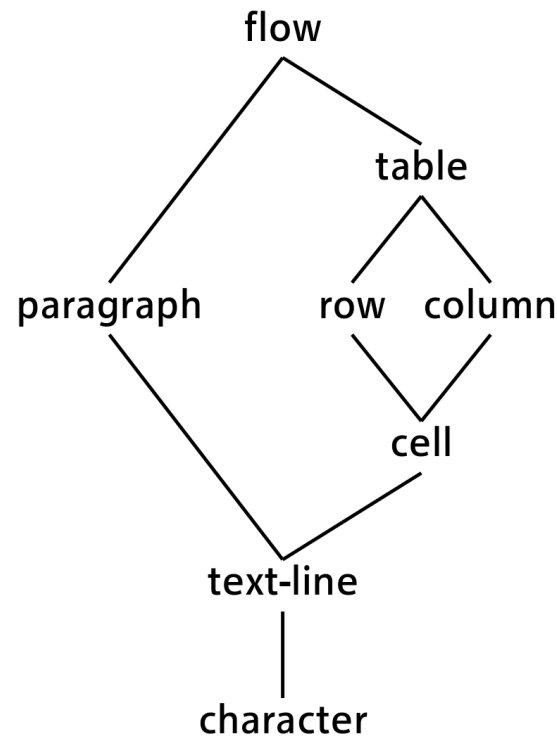
- Variety of approaches, mostly for images:
 - Object recognition (e.g. YOLO)
 - Pixel-based classification
 - Corner/feature detection
 - Pre-segmentation of image
 - Transformation of image
- Excel in simpler table structures and unclean documents
- Black box; difficult to patch; no direct link to PDF content stream
- Require huge dataset for training

Revisiting explainable algorithms



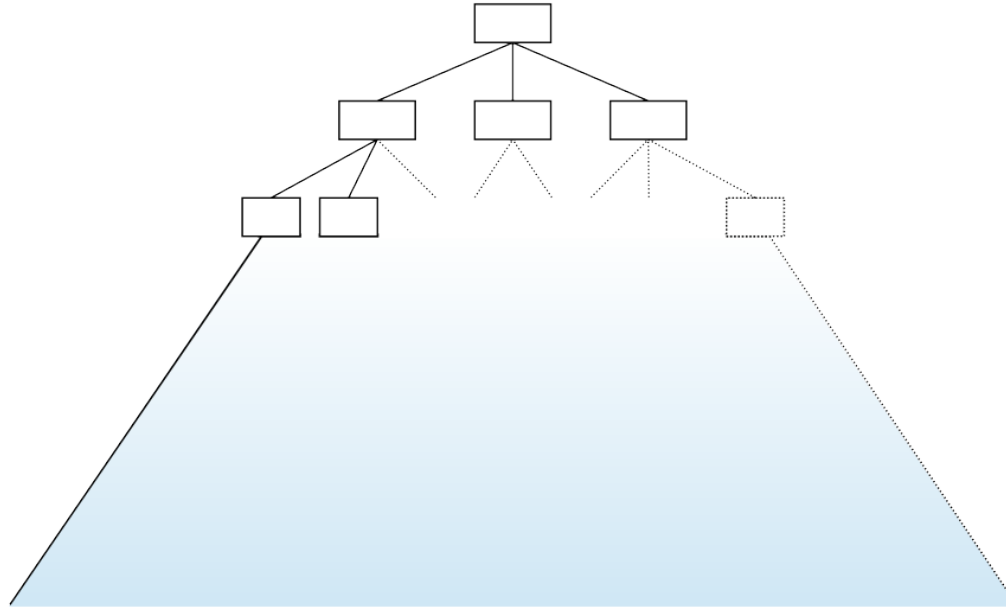
Document analysis as an AI optimization problem

- Two components:
 - Document model with evaluation criteria (moderate)
 - Search through plausible interpretations (hard!)
- Which structures do we want to detect?
- Common to a wide variety of documents



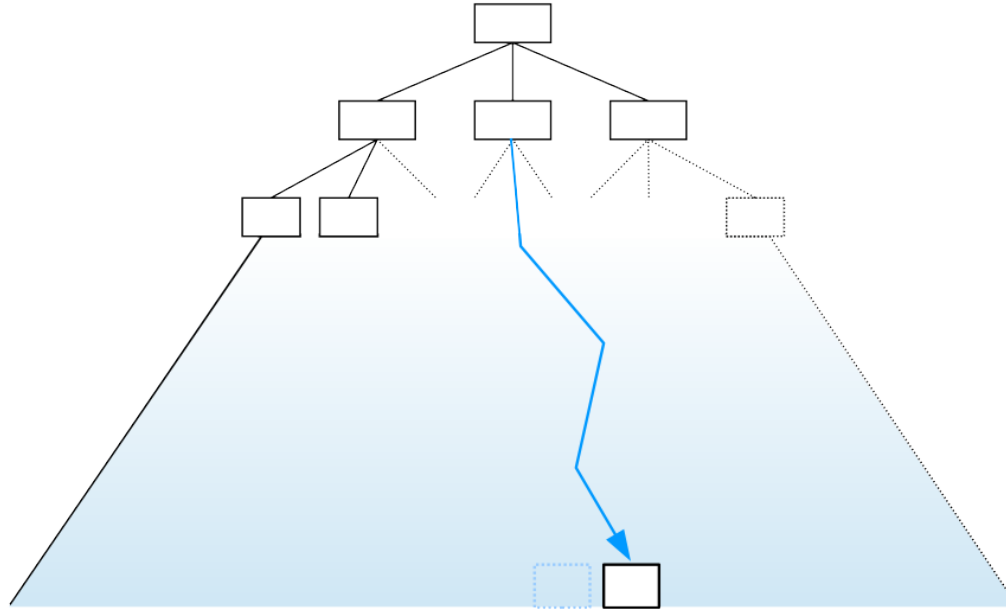
Guided search procedure (1/3)

- Search space



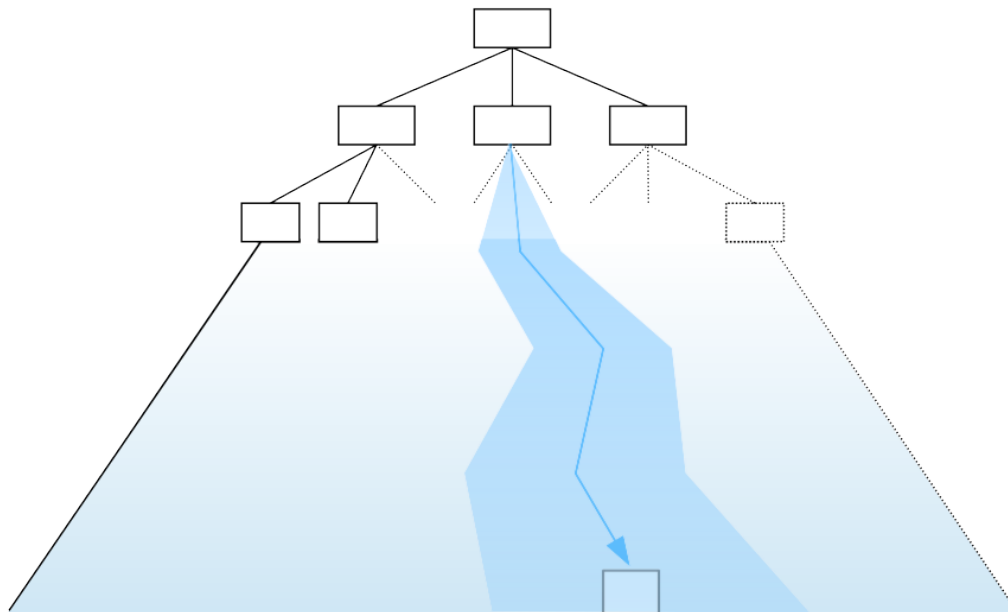
Guided search procedure (2/3)

- Greedy search



Guided search procedure (3/3)

- Guided search



Rule-based vs AI optimization strategies

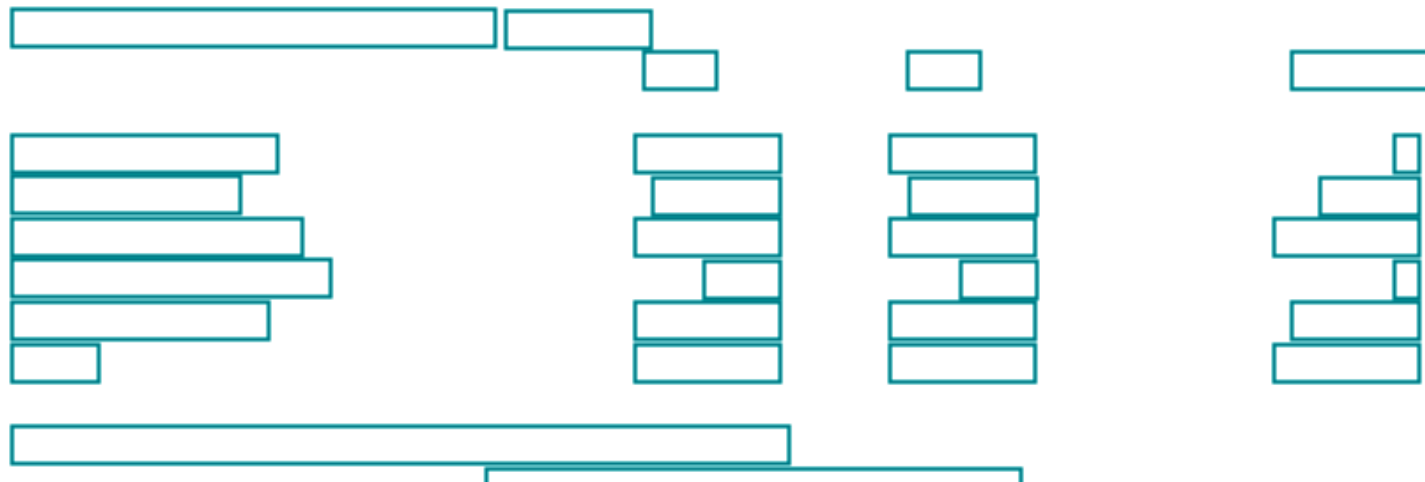
Rule-based

- Limited accuracy due to error propagation
- Easy to implement
- Fast
- Appropriate if source is known

AI optimization

- Good accuracy for all documents supported by model
- Evaluation returns a confidence measure
- Implementation is complex
- Processing time is longer, depending on search strategy

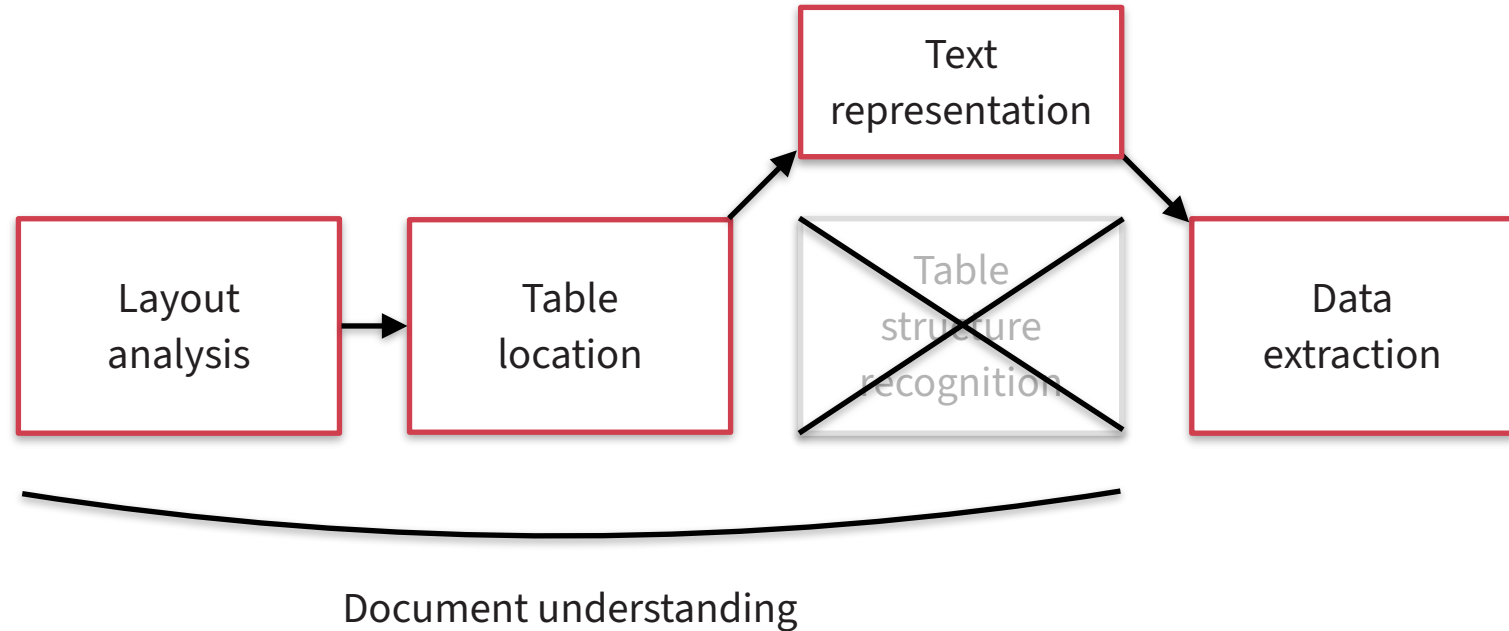
Aside: No domain knowledge



Generative AI



Generative AI option 1



Text representation

Differences with respect to Germany	Portugal	Greece	Spain	Italy	France
Q2 2006 – Q1 2010**	0.33	0.51	0.31	0.18	0.05
Q1 2009 – Q4 2009	-0.01	0.45	0.21	0.18	0.05
Q4 2009	0.17	0.70	0.26	0.09	-0.01
Q1 2010	0.64	0.72	0.56	0.43	0.25



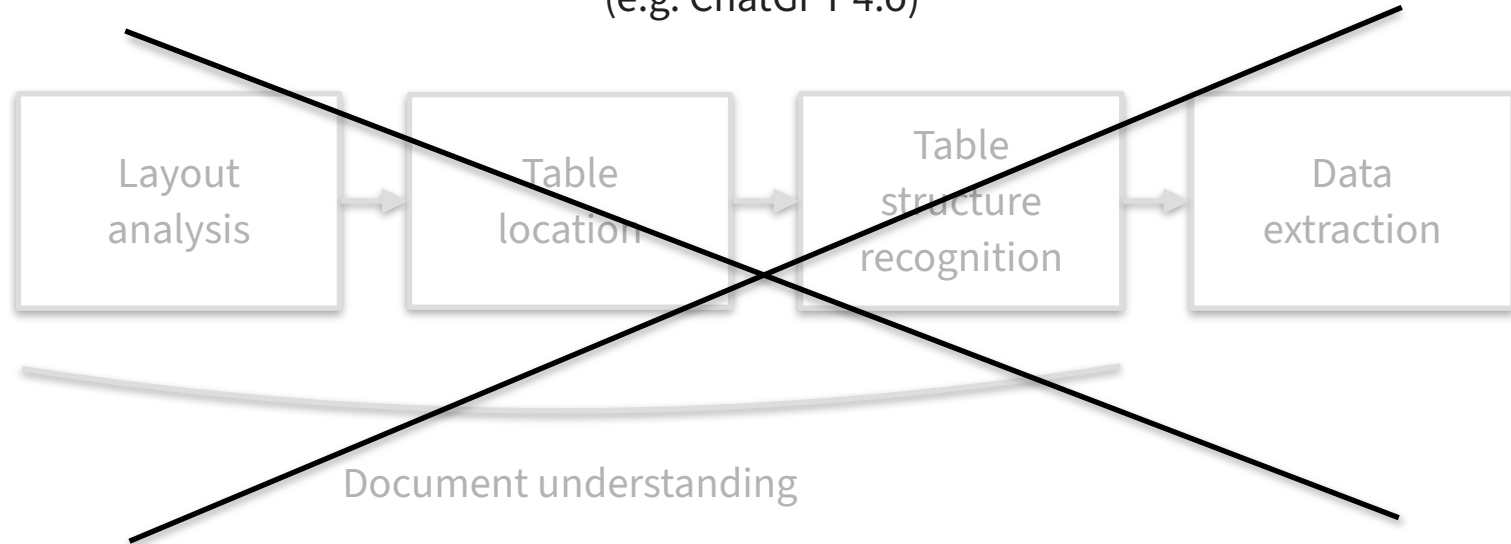
Differences with respect to Germany	Portugal	Greece	Spain	Italy	France
Q2 2006 – Q1 2010**	0.33	0.51	0.31	0.18	0.05
Q1 2009 – Q4 2009	-0.01	0.45	0.21	0.18	0.05
Q4 2009	0.17	0.70	0.26	0.09	-0.01
Q1 2010	0.64	0.72	0.56	0.43	0.25

“According to the following table, what was the figure for Spain in Q4 2009?”

Source: ICDAR 2013 Table Competition Dataset

Generative AI option 2

LLM with vision model
(e.g. ChatGPT 4.o)



Vision model

Differences with respect to Germany	Portugal	Greece	Spain	Italy	France
Q2 2006 – Q1 2010**	0.33	0.51	0.31	0.18	0.05
Q1 2009 – Q4 2009	-0.01	0.45	0.21	0.18	0.05
Q4 2009	0.17	0.70	0.26	0.09	-0.01
Q1 2010	0.64	0.72	0.56	0.43	0.25



“According to the following table, what was the figure for Spain in Q4 2009?”

“Please output the attached table in HTML format.”

Here is the HTML representation of the table shown in your image:

```
html

<table border="1" cellpadding="5" cellspacing="0">
<thead>
<tr>
<th>Differences with respect to Germany</th>
<th>Portugal</th>
<th>Greece</th>
<th>Spain</th>
<th>Italy</th>
<th>France</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q2 2006 – Q1 2010**</td>
<td>0.33</td>
<td>0.51</td>
<td>0.31</td>
<td>0.18</td>
<td>0.05</td>
</tr>
```

Explainable vs generative AI strategies

AI optimization (explainable)

- Good accuracy for all documents supported by model
- Fully explainable result
- Evaluation returns a confidence measure
- Implementation is complex
- Typically requires less processing power than GenAI
- Typically no domain knowledge

Generative AI (RAG with vision model)

- Generally good accuracy for all documents typically found on the Web
- Generally no confidence measures
- Easy to implement
- May require greater processing power than AI search
- May hallucinate; can be combined with other approaches to detect such cases

Combining GenAI with explainable methods

- Resort to generative AI when explainable methods fail to find a suitable result (with sufficient confidence)
 - might be stuck in a local maximum
- Use the explainable method's evaluation component to verify the plausibility of the GenAI result

Thank you

Tamir Hassan
tamir@tamirhassan.com

