

From **Breaking Bad** to **Breaking Data**

Darius Tetsa & Andras Kasa
UCB Pharma S.A.

PHUSE EU Connect 2025



Inspired by **patients**.
Driven by **science**.





Agenda

1. Our Challenge (Breaking the Mindset: PDFs as Data)
2. Our methodology and experimentations
3. Our Results and further adjustments
4. Potential Applications & Use Cases – SAVE Framework
5. Conclusion & Future Work

Quick Poll: Your Experience



Q1: Have you ever tried to find the list of subgroup analyses conducted for a specific study?



Q2: Have you ever struggled to find a similar output or analysis from a past study?



Q3: Would you be interested in getting a summary of all the post-hoc analysis that was done for a given subgroup, by data cut?

Output-Related Findability Questions

Our Challenge

Version & Currency:

version? What is the data cut-off?

studies used

multiple imputation and how?

Analysis
analysis

How do we extract the metadata?



From **Breaking Bad** to **Breaking Data**

When Good Documents Break Down Better



Inspired by **patients.**
Driven by **science.**

*Both pictures were generated with an AI picture generator

Understanding the structure of an output

The Header section: contains deliverable information associated with the TFL.

The Title section: contains the title of the output.

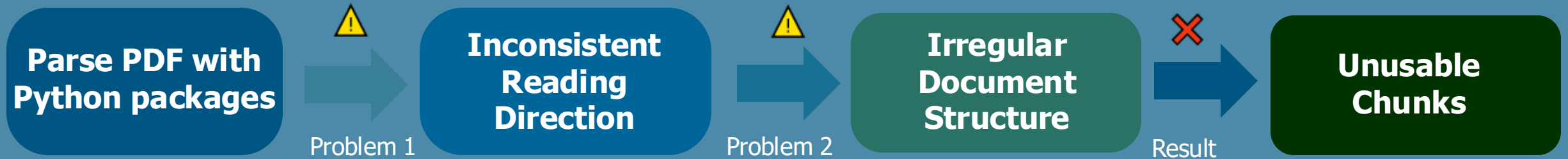
The Body section: contains the actual tables, figures, or listings of interest.

The Footer section: contains information to help facilitate understanding of the body section.

Optionally we may also include a grouping variable for the analysis results at the beginning of the body section of the TFL output.

[illegible]

The Limitation of Simple Chunking Strategies



Coordinate-Based Extraction: Almost There...

Position Instability

- Footer Coordinates vary
- Content length pushes sections up/down (titles)
- "By" statements in tables can be mistaken for titles

Template Fragmentation

- Each vendor uses different templates
- Some acquired products template is different to ours
- Internal formats $\neq$ External formats

Boundary Ambiguity

- Headers and titles run together with no spacing
- No clear Y-coordinates to separate them

What we need

- Semantic understanding
- Vendor/compound agnostic solution
- Understand document structure without supervision

Using Simple Coordinates or Lines to Separate Information

Have you ever seen a person who wears their pants too high?



CONFIDENTIAL
Draft

Integrated Summary of Efficacy

Table 12.1.3.2

Plasma Concentration (ug/ml) by Status and Visit during the Initial Treatment Period - Study Participants

Analysis Set: Pool S1 (Study)

Visit	n	Geometric Mean	Geometric 95% CI	Geometric CV (%)	Median	Mean	Standard Deviation	Minimum	Maximum
BLQ=Below Limit of Quantification, CI=Confidence Interval, CV=Coefficient of Variation, ITP=Initial Treatment Period, LLOQ=lower limit of quantification, Q4W=every four weeks, SD=standard deviation.									
Note: Values BLQ are replaced by value of LLOQ/2 in calculations of means, SDs and CVs; LLOQ/2=0.125ug/mL.									
Note: Means, SDs and CVs are only calculated if at least 2/3 of concentrations and a minimum of 3 datapoints were quantified at the respective timepoint.									
Note: positive only=One or more positive samples, no positive samples during ITP. positive only=One or more positive samples during ITP. positive=One or more positive samples and one and more samples during ITP. negative=all available ADAb positive samples or no more than one missing or inconclusive and all other samples are NAb negative during ITP.									
Note: n excludes samples without PK results and samples with significant deviations that impact the PK concentration at the respective visit.									

Program: Page 1 of 6

It is clear that coordinates can give you totally wrong directions, not just because of this example, but because titles can be broken into 4 or 5 lines, and sometimes headers can be moved upwards.

Unsupervised learning

We tested hierarchical clustering to automatically discover document structure and assist with annotation, but the approach produced inconsistent and unreliable segmentation

- Too granular (over-segmentation)
- Not granular enough (under-segmentation)

Cluster 4

CONFIDENTIAL
Draft

Table 4.1.1

Cluster 3

	Placebo (N=211)	Active (N=211)
Wakpoeder Rate % (95% CI)		
n/Nsub (%)		
Odds Ratio vs Placebo		2.48
97.5% CI for Odds Ratio		
nominal p-value		<0.001
Difference vs Placebo		20.6
95% CI for difference		

Cluster 1

The model [redacted] and baseline Weight Band category [redacted]

Program: [redacted]

Page 1 of 1

Supervised learning – Computer Vision

You live only once, do not waste your time

YOLO is a deep learning model that was first introduced in 2015. The abbreviation stands for **"You Only Look Once."**

Compared to its predecessors, this model requires only one forward propagation pass through the neural network to make predictions, while other models rely on region proposal techniques requiring thousands of iterations

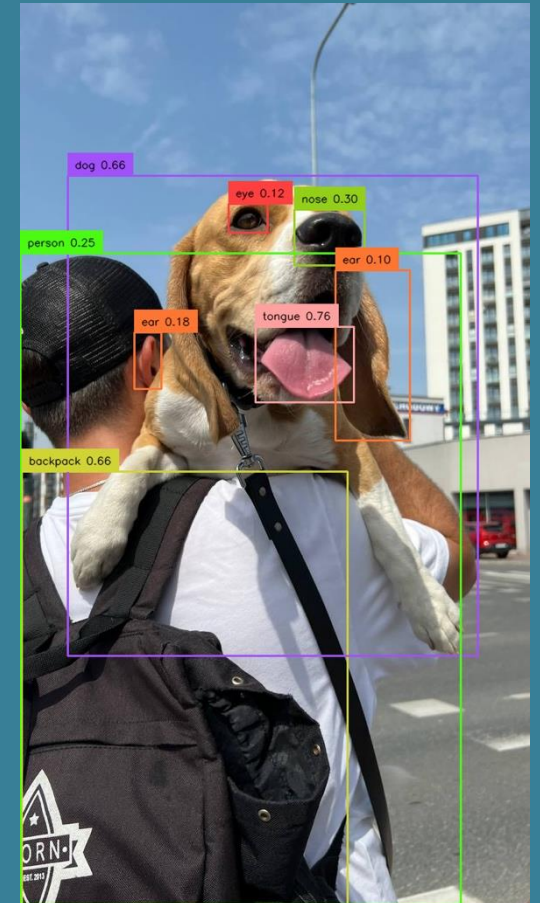
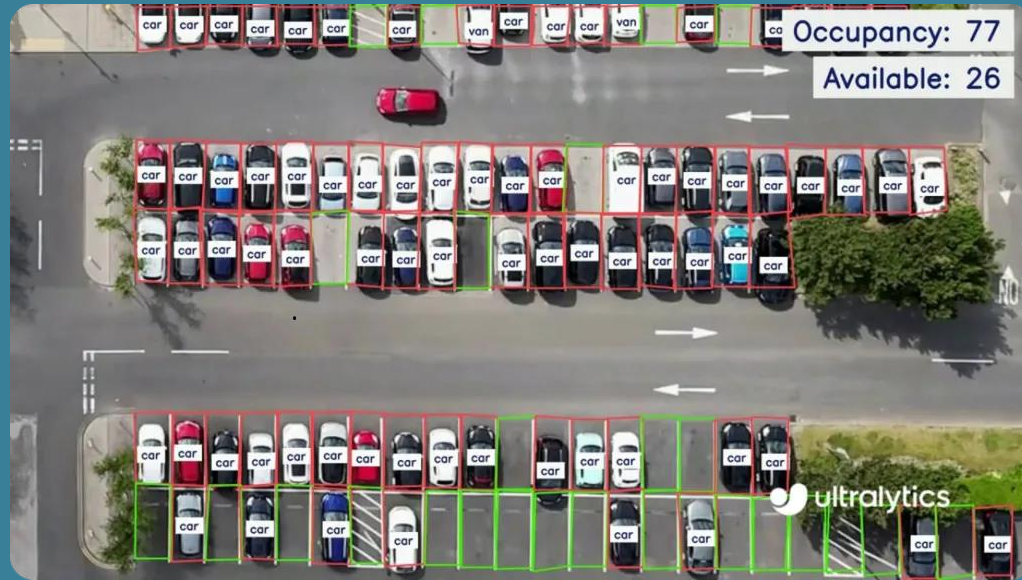
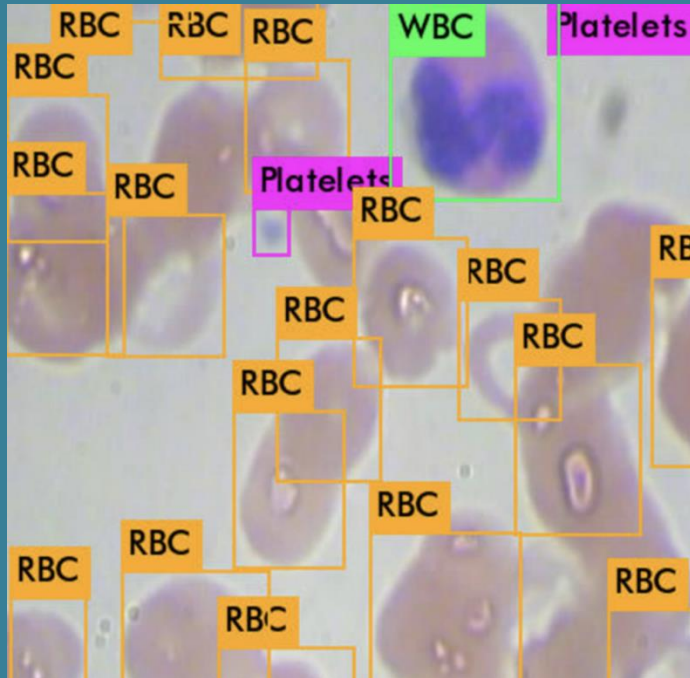
It is significantly faster and more efficient than classical convolutional networks, making it ideal for real-time applications

As output, it provides both the name of potential objects and a confidence score indicating the probability of correct identification

YOLO has diverse industrial applications, including:

- Factory automation for counting and quality control of items
- Object and person detection across various environments
- Vehicle speed measurement and traffic monitoring
- Security and surveillance systems
- Medical imaging analysis

Examples of YOLO



Why Yolo?

Easy to tweak

YOLO demonstrates excellent transfer learning capabilities - while not specifically trained on documents initially, its robust architecture can effectively adapt to learn document structures and layouts

The model's efficiency in transfer learning means:

- Only a few hundred training images can be sufficient for specific tasks
- It leverages pre-trained knowledge from millions of diverse images
- Its billions of parameters can be fine-tuned for specialized applications
- Quick adaptation to new domains without extensive retraining

Collect, annotate Tables/Listings

We annotated and trained on 4 classes: header, titles, body, and footer.

One can observe that each class has typical features, such as headers typically forming a rectangle since the left side is left-aligned while the right side is right-aligned.

 Inspired by **patients**.
Driven by **science**.

Header	
TNT GROUP	CONFIDENTIAL Draft
Title	Table 2.2
Plasma Concentrations of [REDACTED]	[REDACTED]
Body	
[REDACTED]	
	Time [h]
Subject	0 0.5 1 2 3 4 6 8 12 24 48
[REDACTED]	
[REDACTED]	
[REDACTED]	
[REDACTED]	
[REDACTED]	
n	6
[REDACTED]	
[REDACTED]	
[REDACTED]	
[REDACTED]	
[REDACTED]	
[REDACTED]	
[REDACTED]	BLQ
[REDACTED]	BLQ
[REDACTED]	BLQ

Our training - Testing on new data

- **Training Parameters:**

- 50 epochs of training
- Successfully identified main document objects
- Model showed robust recognition capabilities

- **Additional Details:**

- Training set: 736 annotated documents
- Validation set: 210 images
- Test set: 106 independent files
- Data augmentation: copy_paste=0.1, mixup=0.2

- **Overall Model Performance**

- Model caught 99.6% of all objects (high recall: 0.996)
- About 99 out of 100 detections were correct (precision: 0.991)
- Overall effectiveness score of 99% (F1 Score: 0.99)
- mAP@50: 0.994 (excellent boundary accuracy)

- **Performance by Category**

- All four categories achieved perfect or near-perfect accuracy:
- Header, Title, Body, Footer: 100% classification accuracy

YOLO achieves 100% section detection accuracy

BUT Perfect box detection ≠ Perfect text extraction

header: 0.86

CONFIDENTIAL
Final

title: 0.88

Table 6.1.1
Analysis Set: Safety Set iv

body: 0.93

Age Cohort Visit (Descriptor)	Observed Result						Change From Baseline* Result					
	n	Mean	SD	Median	Min	Max	n	Mean	SD	Median	Min	Max
>=1 month - <8 years (N=5)												
Baseline*												
Visit 1a, Screening (Day -1)												
Visit 1b, Baseline (Day 1)												
Final Visit												
>=8 - <17 years (N=16)												
Baseline*												
Visit 1a, Screening (Day -1)												
Visit 1b, Baseline (Day 1)												
Final Visit												

footer: 0.94

Reference: Listing 8.1

Program: t_lb611, 2019-08-28 at 15:56

Page 208 of 784

Page 73 of 150

CONFIDENTIAL
Final

header=0.86

Table 6.1.1

title=body

Age Cohort Visit (Descriptor)	n	Observed Result					n	Change From Baseline* Result				
		Mean	SD	Median	Min	Max		Mean	SD	Median	Min	Max
>=1 month - <8 years (N=5)												
Baseline*												
Visit 1a, Screening (Day -1)												
Visit 1b, Baseline (Day 1)												
Final Visit												
>=8 - <17 years (N=16)												
Baseline*												
Visit 1a, Screening (Day -1)												
Visit 1b, Baseline (Day 1)												
Final Visit												

body=footer

lacosamide, SD=standard deviation.

Reference: Listing 8.1

Page 208 of 784

Page 73 of 150

Consensus Voting - SODA Method

Sample

- First 5 pages of each output
- 6 random pages
- Last page

Note: Maximum 12 pages

Observe

- Measure 3 boundaries per page
- Header-Title
- Title-Body
- Body-Footer

Decide

- Collect all measurements
- Create a consensus of the texts per classis

Apply

- Single boundary set
- Applied to all pages

For deciding section delimitation in multi-pages outputs

Extracting Table Content

Why PyMuPDF?

- Dictionary based `get_text("dict")`
- Preserves spatial relationship

How it works?

- Groups by Y-coordinate (identifies rows)
- Sorts by X-coordinate (maintains columns)
- Uses whitespaces+linebreaks
- LLM compatible
Output: Modern LLMs interpret structure for Q&A

~~Figure 5. EVDS Symptom Onset Group~~ <=2 years of age (N=21)

Baseline Health State	Year 1 Health State						Death n (%)
	Early Ambulatory n (%)	Late Ambulatory n (%)	Non Ambulatory n (%)	Ambulatory NIV 4+ n (%)	Non Ambulatory NIV 4+ n (%)	Non Ambulatory IV 8+ n (%)	
Early Ambulatory							
Late Ambulatory							
Non Ambulatory							
Ambulatory NIV 4+							
Non Ambulatory NIV 4+							
Non Ambulatory IV 8+							
Death							

~~Figure 6. EVDS Symptom Onset Group~~ <= 2 years of age (N = 21)

Year 1 Health State							
Baseline Health State	Early Ambulatory	Late Ambulatory	Non Ambulatory	Ambulatory NIV 4+	Non Ambulatory NIV 4+	Non Ambulatory IV 8+	Death
Early Ambulatory							
Late Ambulatory							
Non Ambulatory							
Ambulatory NIV 4+							
Non Ambulatory NIV 4+							
Non Ambulatory IV 8+							
Death							

Potential Applications & Use Case: The SAVE Framework

S A V E

Searchable Knowledge Base

Build a
comprehensive
repository of past
studies, methods,
and
implementations

Automated Reasoning

Enable smarter AI
systems using
historical patterns
and proven
approaches

Validation & Quality Control

Extract and
verify statistical
results,
population
numbers, and
consistency

Extractable Metadata

Capture
metadata from
outputs
generated by
external
vendors

Conclusion & Future Work

What we achieved? ✓

- 99% Classification Accuracy
- Consensus Voting & Geometric Boundaries (100% agreement)
- Table/Listing content extraction with PyMuPDF
- LLM-compatible Output

Where are we heading? →

- Extend to Figures
- NLP Integration
- Knowledge Graph
- Continuous Learning
- Experimenting with multimodal models

Our YOLO repository:



Thanks. Any feedback?

Please send us an email:
Darius.Tetsa@ucb.com
Andras.Kasa@ucb.com



Inspired by **patients.**
Driven by **science.**

Proprietary and Confidential Property of UCB

