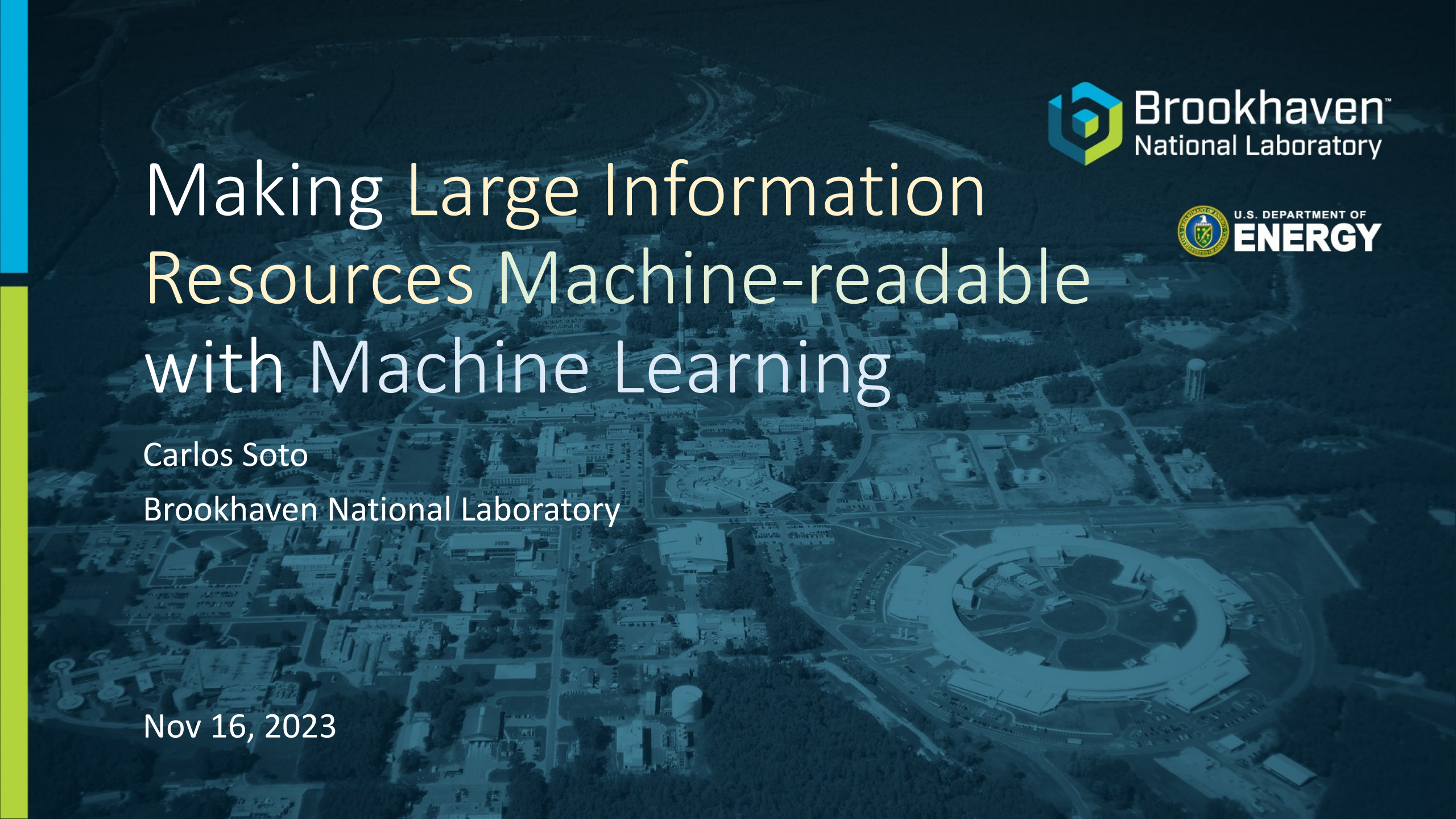


Making Large Information Resources Machine-readable with Machine Learning

Carlos Soto

Brookhaven National Laboratory

Nov 16, 2023

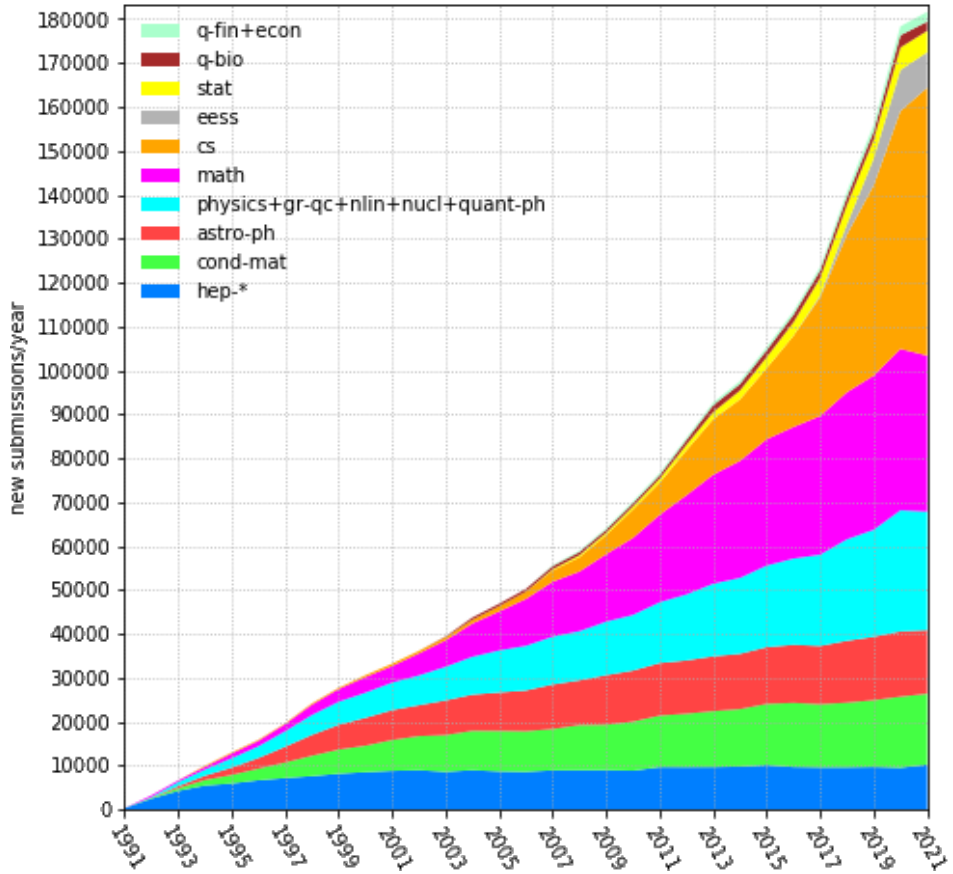
Outline

- Information at Scale
- Successes with Text and Language
- Documents
- Tables
- Figures & Charts
- A Vision for ML-driven Knowledge Curation


```
{
    "orders": [
        {
            "orderno": "748745375",
            "date": "June 30, 2088 1:54:23 AM",
            "trackingno": "TN0039291",
            "custid": "11045",
            "customer": [
                {
                    "custid": "11045",
                    "fname": "Sue",
                    "lname": "Hatfield",
                    "address": "1409 Silver Street",
                    "city": "Ashland",
                    "state": "NE",
                    "zip": "68003"
                }
            ]
        }
    ]
}
```

Information at Scale

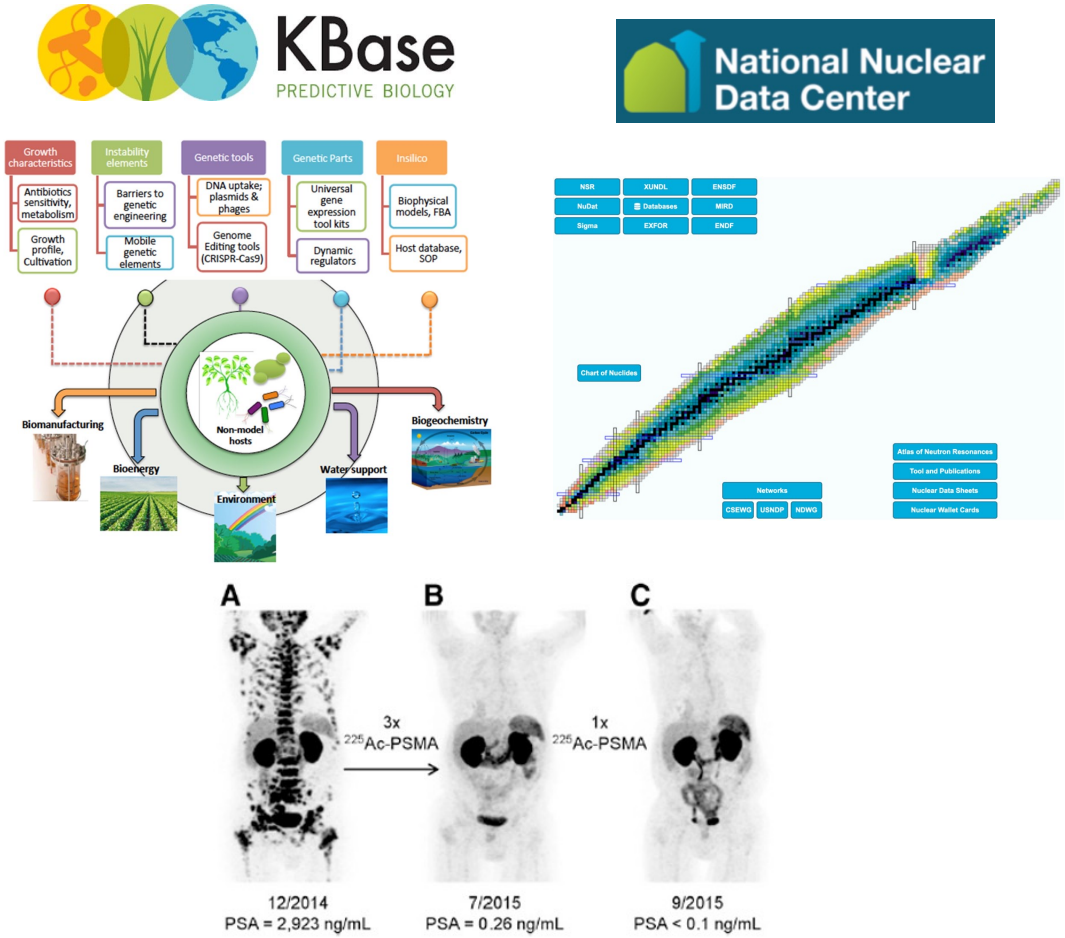
- Many missions and efforts require analyzing vast information resources, much of whose contents are **not machine-readable**. These resources:
  - Inform the design, targets, and parameters of **new experiments**
  - Serve as **context** for interpreting and comparing results
  - Offer opportunity to perform **large-scale analyses**
- **Massive scale** limits effective use, inhibits full potential
  - e.g., millions of existing articles, publication rates accelerating
  - **Impossible to manually survey** all relevant works



Submission rates to arXiv

Some of my Motivating Applications

- Biology
  - Genetic toolkit and protein-protein interaction (PPI) **discovery**
- Chemistry
  - Parameterization of isotopes separations **experiments**
- Nuclear Physics
  - **Curation** of experimental nuclide structure quantities into authoritative resource
- Commonalities:
  - Overcome **highly manual** workflows
  - Enable observation of **emergent patterns**

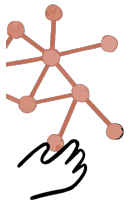


Top-left: BER KBase project components and applications; *top-right*: visualization of BNL NNDC's managed nuclide data; *bottom*: highly successful prostate cancer clinical trial with Actinium-225, one of the isotopes of interest at BNL's Isotope Research and Production Department.

Brookhaven National Laboratory

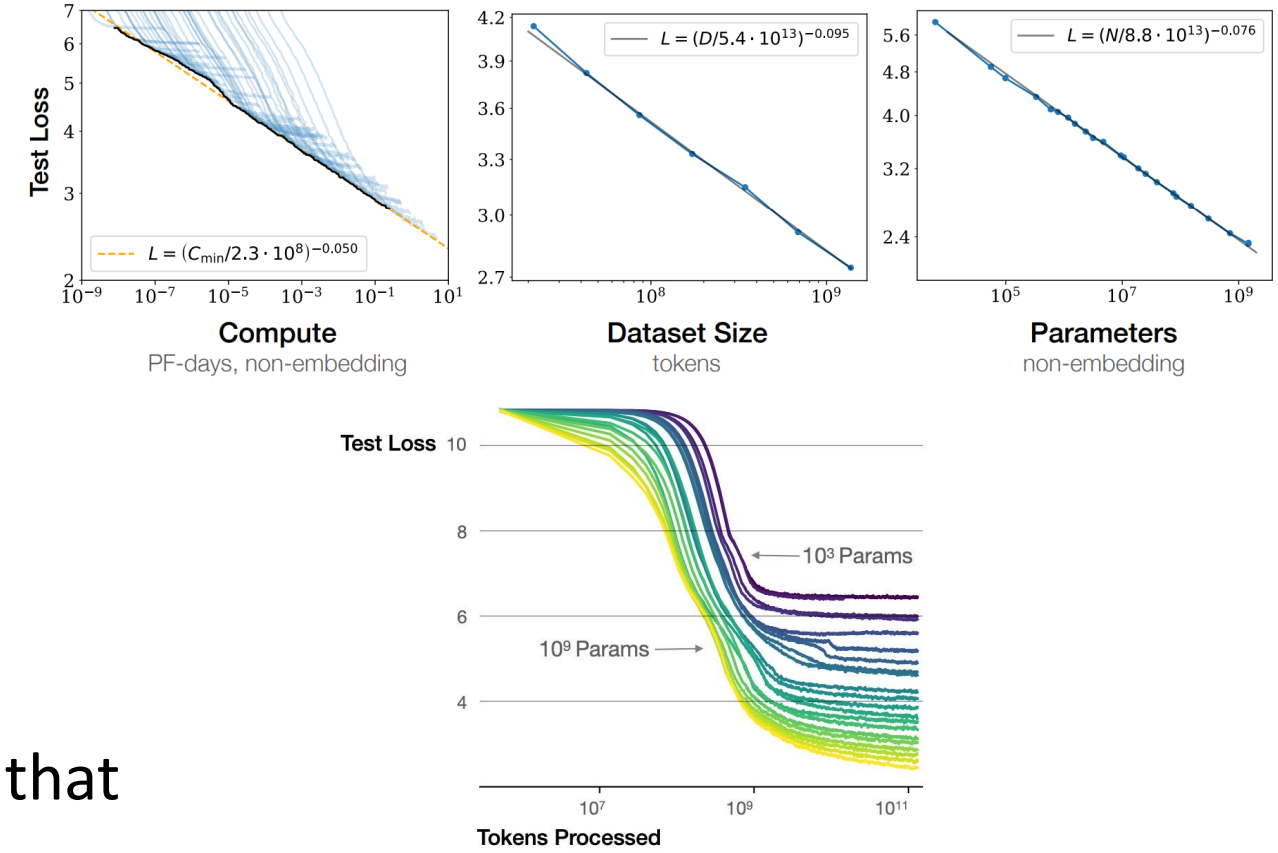
A highly visible success story: Natural Language Processing

- Large Language Models (**LLMs**) have demonstrated great **knowledge retrieval** capabilities; kicked off new AI race
  - Transformer architecture (2017) enabled scale
  - First LLMs in 2018 (GPT-1, BERT)
  - Public attention in 2022 (ChatGPT)
- LLMs are pretrained with simple language modeling objective using **vast text corpora**
  - Limited to *machine readable text*..



Why Scale Matters for NLP

- LLM performance has power-law relationship with each of:
  - Compute time
  - **Training data (# tokens)**
  - LLM parameters
- Similar properties hold for other modalities
  - e.g., “Foundation models”
- Exploiting these properties requires that training data is machine readable
  - *Web text is easy, other sources are not*



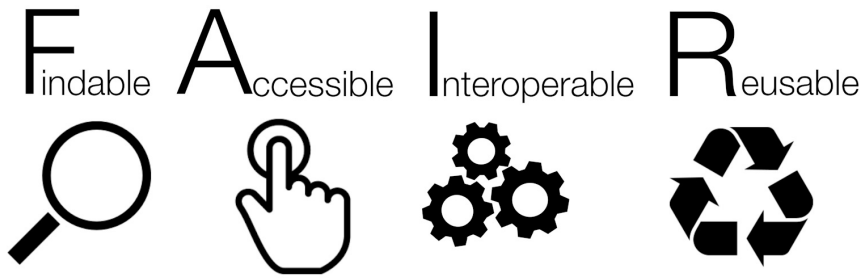
Top: Chinchilla scaling laws[1] showing power-law relationships (lower test loss is better, note that plots are log-log); *bottom*: GPT-3 experiments[2] showing sudden drop in loss (i.e., performance jump) after sufficient training data.

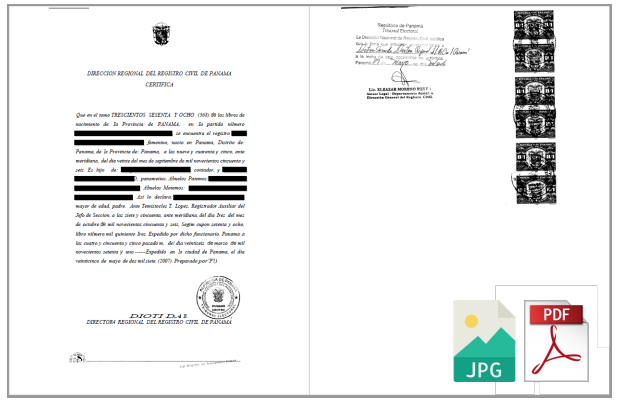
Brookhaven National Laboratory

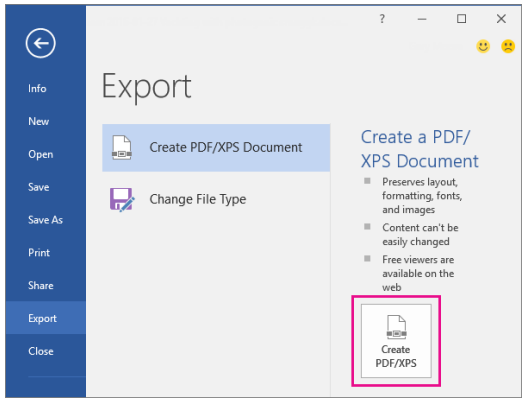
1. Jordan Hoffman, et al. (DeepMind) “Training Compute-Optimal Large Language Models”, arXiv:2203.15556, 2022.
2. Jared Kaplan, et al. (OpenAI) “Scaling Laws for Neural Language Models”, *arXiv:2001.08361*, 2020.

Documents as scans and PDFs

- Increasingly common to encourage or mandate open data & machine-readable documents
  - These standards are still not the norm
  - *Huge historical archives*
- Scanned documents are not trivial to process
  - Naïve OCR enables text search, but not organization; **misses other contents**
- Even born-digital PDFs are not actually machine readable..



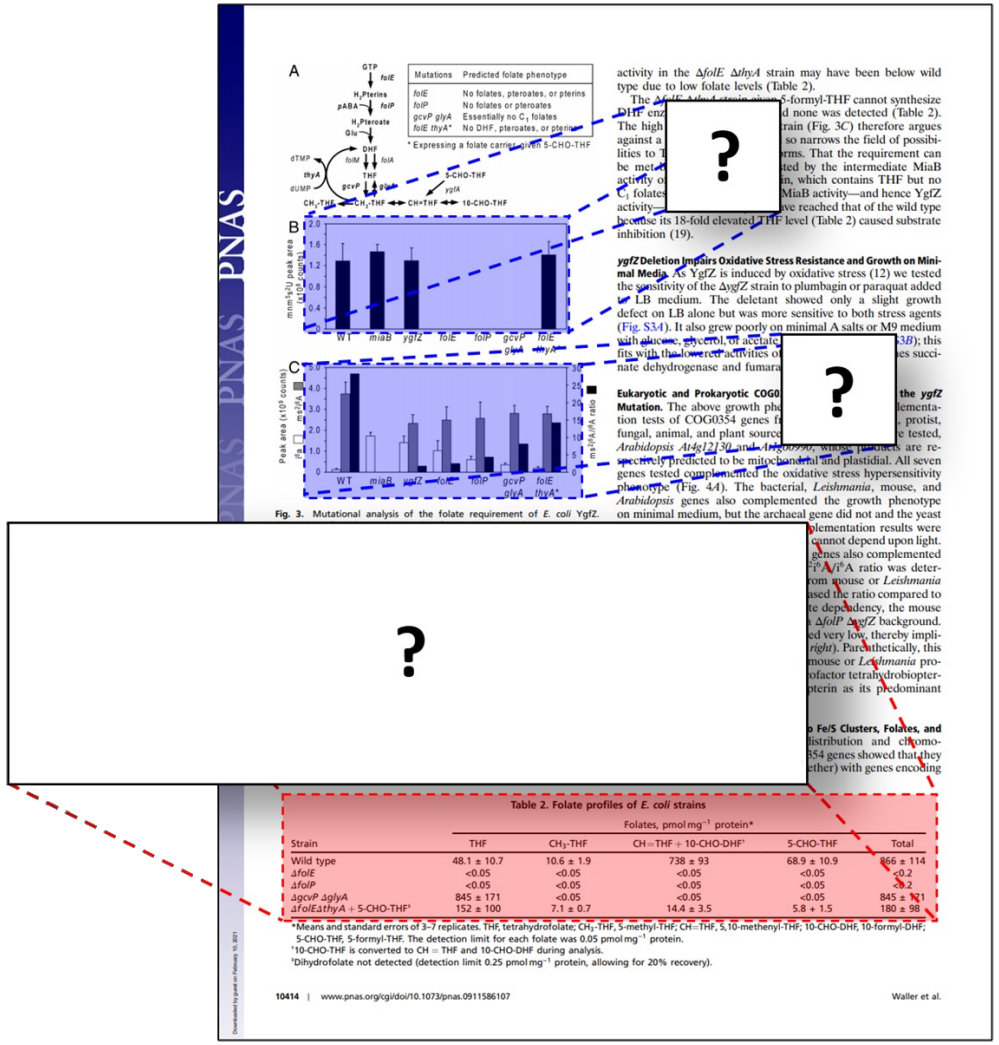




Top: FAIR data principles are gaining usage but remain far from common; *bottom-left*: scanned documents (as images or PDFs) are significant component of many archives; *bottom-right*: digitally sourced PDFs are not a silver bullet.

PDFs

- Algorithmically extracting information from PDFs is *exactly as limiting as from OCR scans*
- Current PDF extraction tools
  - Raw text, parsed **heuristically**
    - Ordering & structure not guaranteed
  - Spurious **artifacts are common** (headers, line numbers, captions, etc.)
    - Impedes text processing (and LLM training)
  - Tables are treated as unstructured text
    - Structure inferred with heuristic rules
  - Figures and **charts are generally ignored**
- **Extraction errors** impair scalable information retrieval & analysis

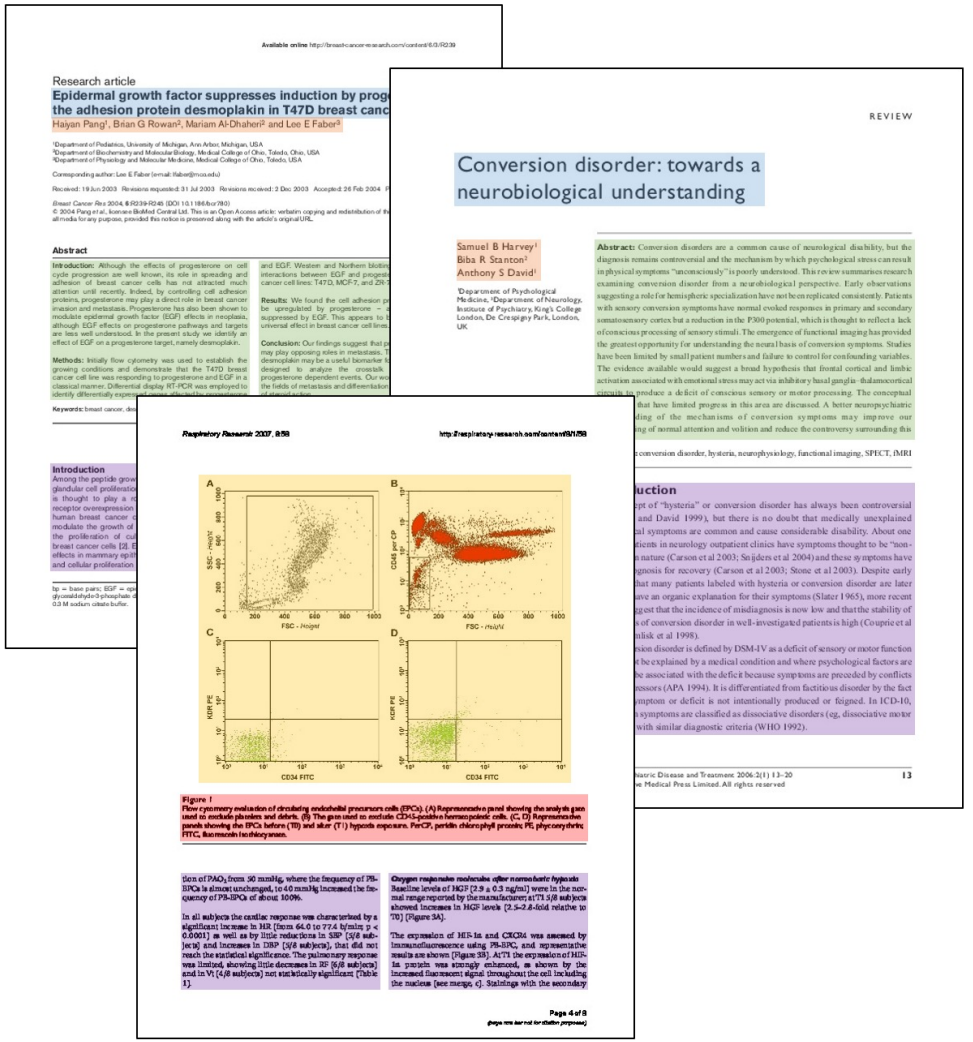


Brookhaven National Laboratory

Document Layout Analysis

To **extract contents of different types** (text, tables, figures, captions, etc.) documents must be reliably and automatically segmented.

- Large portion of information and knowledge in documents takes **non-text form**
- Article PDF files have **highly varied underlying formatting** (even among digital productions)
- Most extraction tools rely on analyzing document *metadata* and markup; results are often very noisy

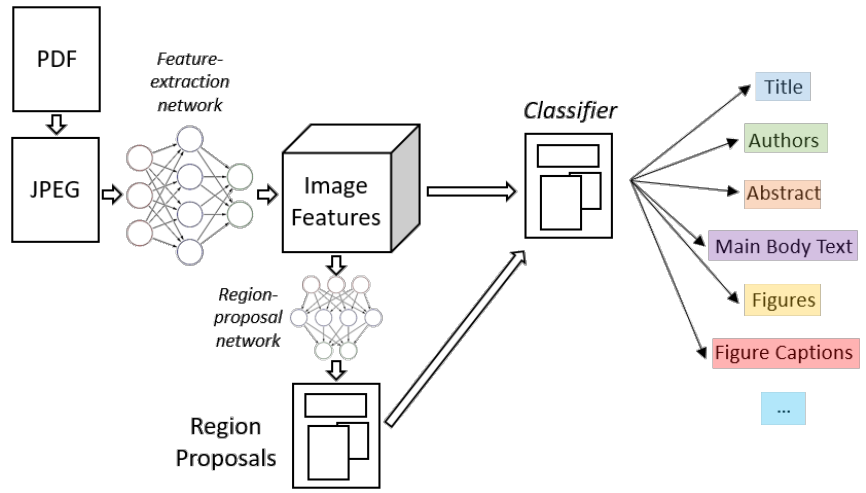


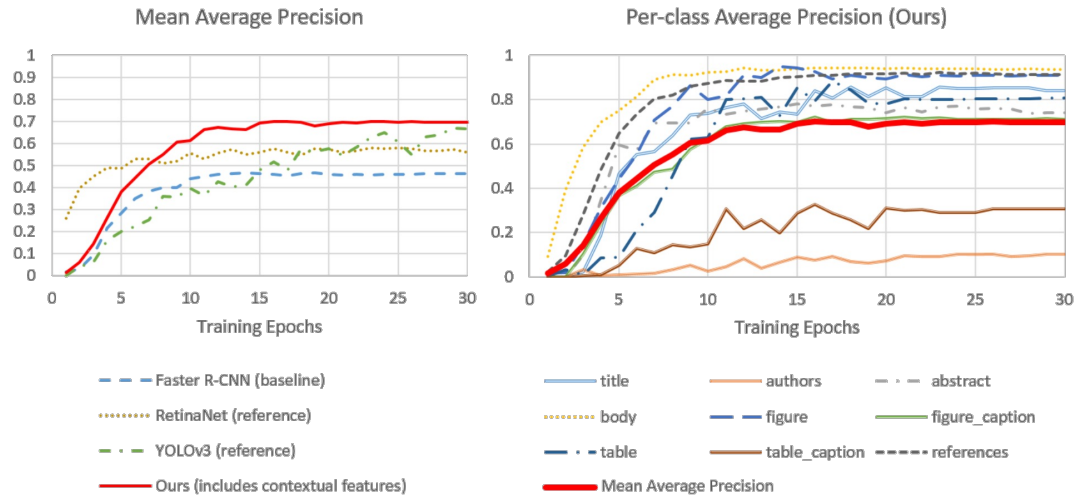
Brookhaven National Laboratory

Document Layout Analysis

Our ML approach[1] is **purely vision-based**

- We adapt an object detection method to exploit the structured nature of scientific articles
- This approach enables processing **tables and figures**, which contain valuable information, yet are inaccessible to most methods
- Our method achieves **94%+ detection precision** on full-text articles, with **14x faster** document processing than prior methods





Brookhaven National Laboratory

1. Carlos X. Soto and Shinjae Yoo, "Visual Detection with Context for Document Layout Analysis." *Empirical Methods in Natural Language Processing* (EMNLP), 2019.

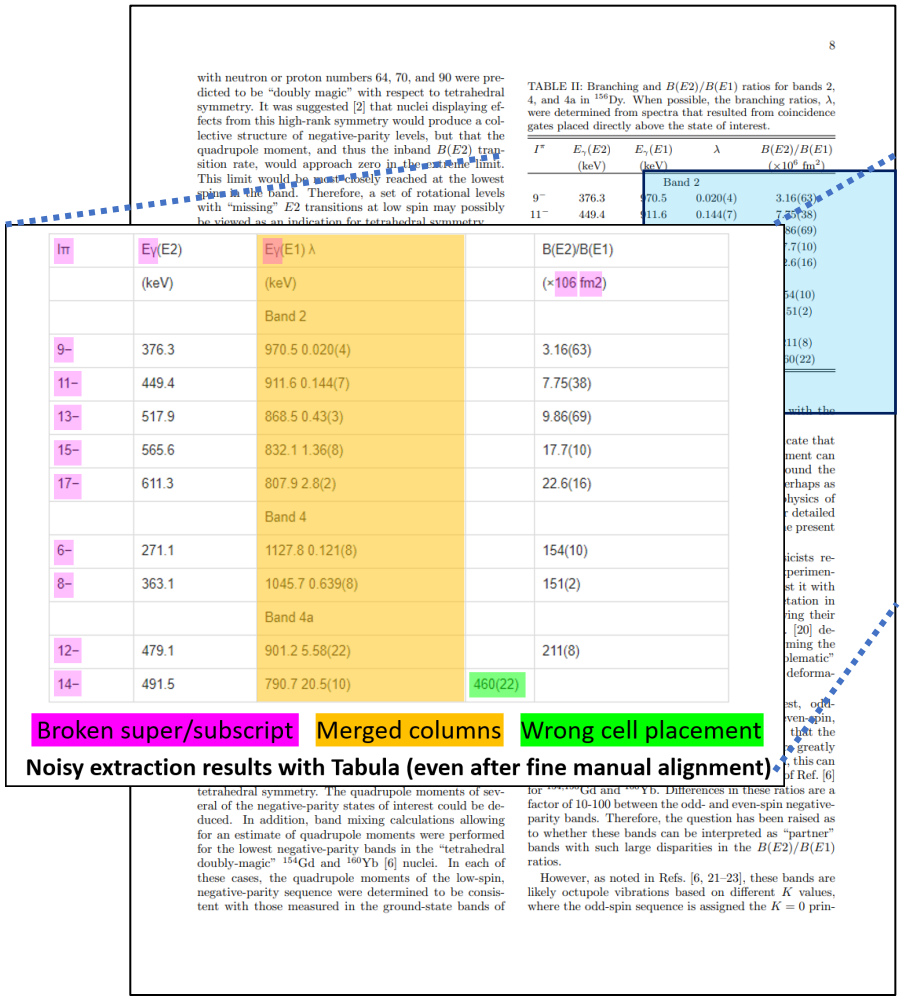
Automated Table Extraction

Tables are theoretically excellent data containers for machine processing, however **PDF-embedded tables are often very challenging to extract**.

- Standard extraction tools (e.g., *Tabula*) require manual table selection, produce many **extraction errors** (*sample with annotated errors at right*)
- Poor performance and manual intervention make such methods **unscalable**

We pursued two approaches to ML-based table extraction: mixed modality, and image-only

- Mixed modality approach used pipeline of adapted and custom models, used *image and text* data
- Achieved excellent table and cell detection (**98.8** and **96.0 F1**, respectively); but poor *structure prediction*

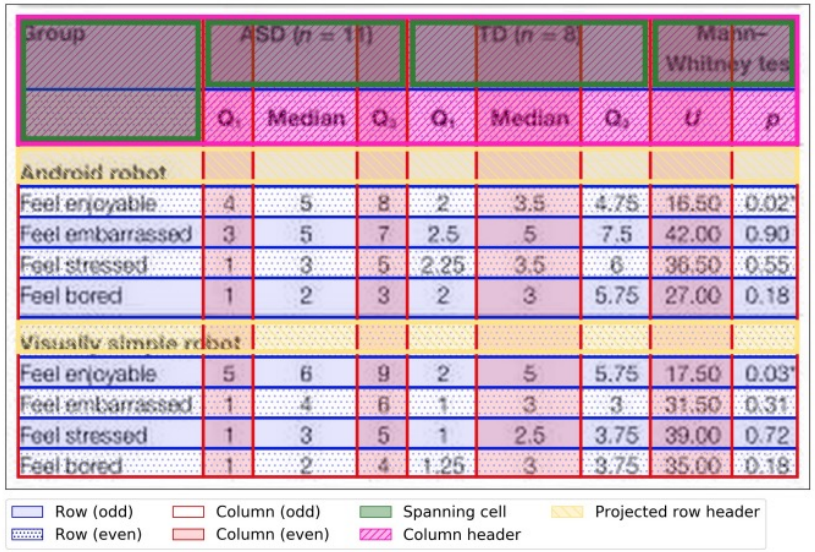


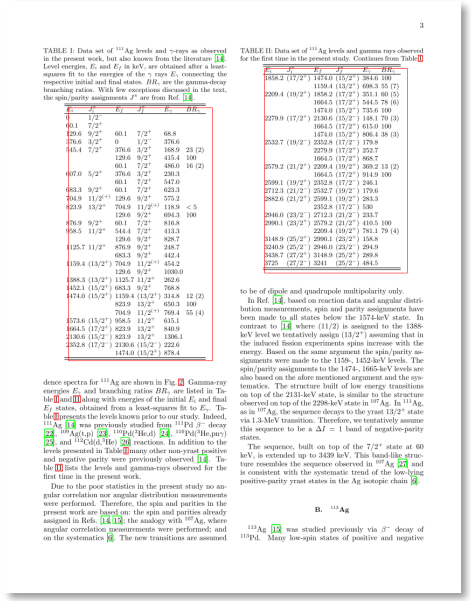
Brookhaven National Laboratory

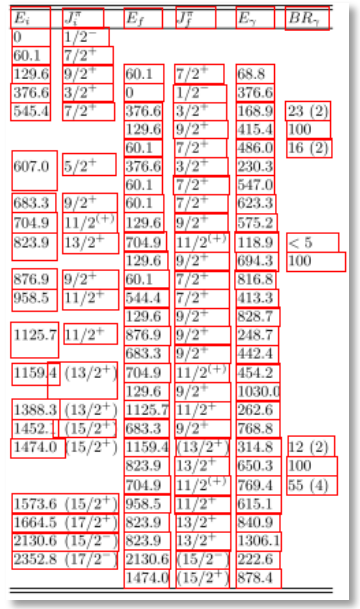
Automated Table Extraction

Image-only approach uses a Transformer-based object detection model (DETR[2])

- Treats structure prediction task as a specialized image task
  - Previous approach was sequence to sequence text prediction (with LSTM or Transformer)
- Pretrained on ~1M machine-annotated tables
  - Fine-tuned on tables from ~100 LaTeX-PDF document pairs
- Improves structure prediction from **78.1 to 98.5** (*grid table similarity metric*)







1. Nicolas Carion, et al. "End-to-end object detection with transformers." *European Conference on Computer Vision* (ECCV), 2020.

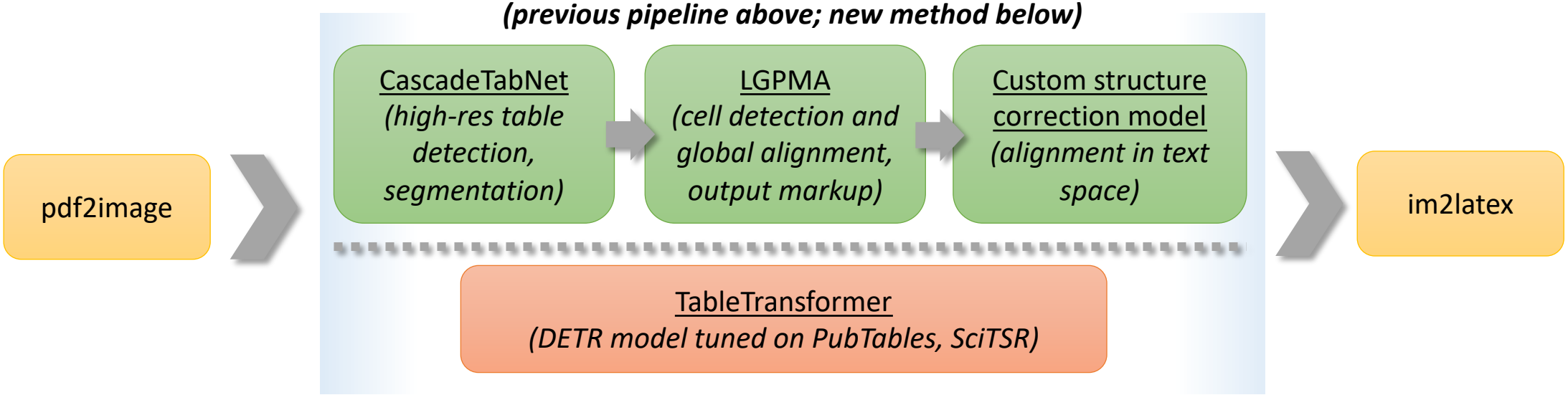
Brookhaven National Laboratory

Automated Table Extraction

PDF pre-processing

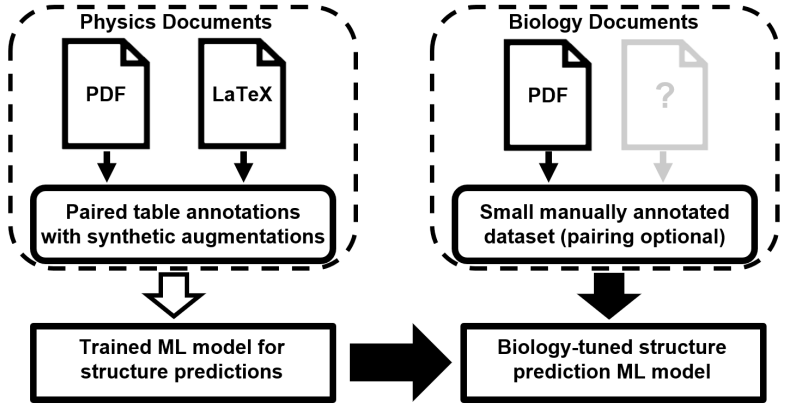
ML-based table detection, segmentation, and structure parsing
(previous pipeline above; new method below)

OCR, post-processing



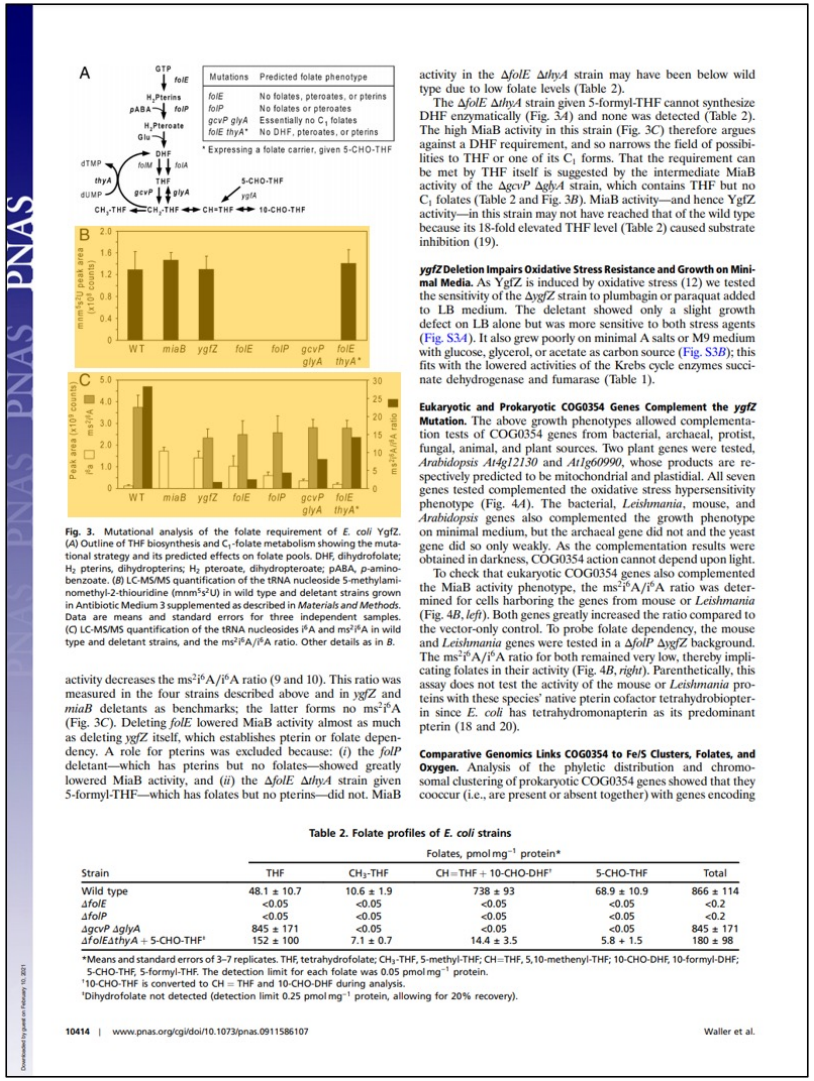
Structure prediction is a key challenge, particularly for domains without structure-resolving annotation pairs (e.g., LaTeX source)

- Addressable to some degree with domain adaptation



Reverse-engineering Data Charts

- Automatically extracting data chart contents could **enable new large-scale analysis** and knowledge discovery tools
  - Charts/plots are currently **inaccessible to *automated* document processing**
- Identifying charts in articles is straightforward
  - Can use whole document layout analysis tool, or simple object detection
  - **>95% *mAP*** with standard object detection models (e.g., *YOLOv3*[3])
- Imperfect alignment, need to oversample; so **value extraction must be robust to artifacts**



activity in the Δ*folE* Δ*thyA* strain may have been below wild type due to low folate levels (Table 2).

The Δ*folE* Δ*thyA* strain given 5-formyl-THF cannot synthesize DHF enzymatically (Fig. 3*A*) and none was detected (Table 2). The high MiaB activity in this strain (Fig. 3*C*) therefore argues against a DHF requirement, and so narrows the field of possibilities to THF or one of its C_1 forms. That the requirement can be met by THF itself is suggested by the intermediate MiaB activity of the Δ*gcvP* Δ*glyA* strain, which contains THF but no C_1 folates (Table 2 and Fig. 3*B*). MiaB activity—and hence YgfZ activity—in this strain may not have reached that of the wild type because its 18-fold elevated THF level (Table 2) caused substrate inhibition (19).

Table 2. Folate profiles of *E. coli* strains

Strain	Folates, pmol mg^{-1} protein*				
	THF	CH_3-THF	CH=THF + 10-CHO-DHF†	5-CHO-THF	Total
Wild type	48.1 ± 10.7	10.6 ± 1.9	738 ± 93	68.9 ± 10.9	866 ± 114
Δ*folE*	<0.05	<0.05	<0.05	<0.05	<0.2
Δ*folP*	<0.05	<0.05	<0.05	<0.05	<0.2
Δ*gcvP* Δ*glyA*	845 ± 171	<0.05	<0.05	<0.05	845 ± 171
Δ*folE*Δ*thyA* + 5-CHO-THF‡	152 ± 100	7.1 ± 0.7	14.4 ± 3.5	5.8 + 1.5	180 ± 98

*Means and standard errors of 3–7 replicates. THF, tetrahydrofolate; CH_3-THF, 5-methyl-THF; CH=THF, 5,10-methenyl-THF; 10-CHO-DHF, 10-formyl-DHF; 5-CHO-THF, 5-formyl-THF. The detection limit for each folate was 0.05 pmol mg^{-1} protein.

†10-CHO-THF is converted to CH = THF and 10-CHO-DHF during analysis.

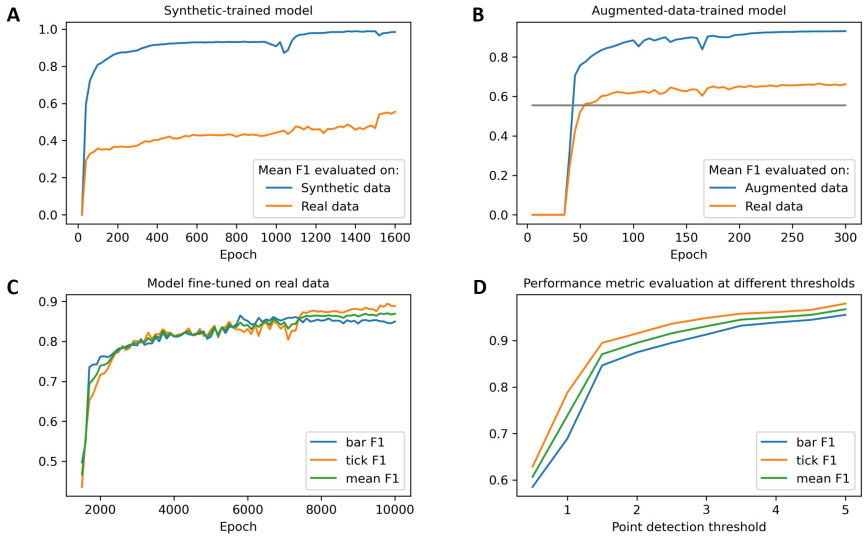
‡Dihydrofolate not detected (detection limit 0.25 pmol mg^{-1} protein, allowing for 20% recovery).

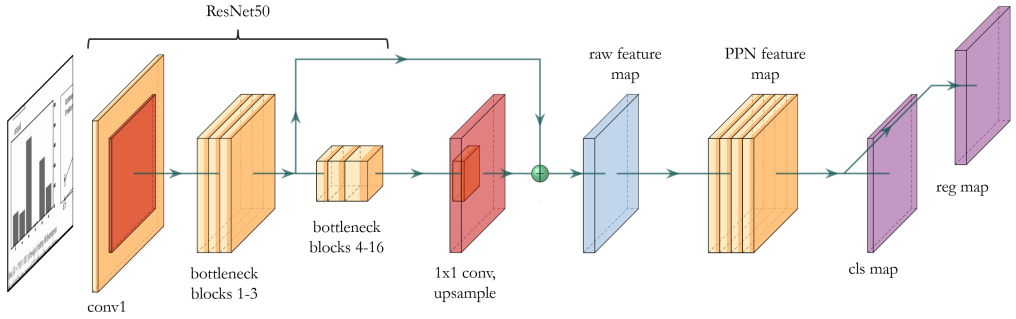
Brookhaven National Laboratory

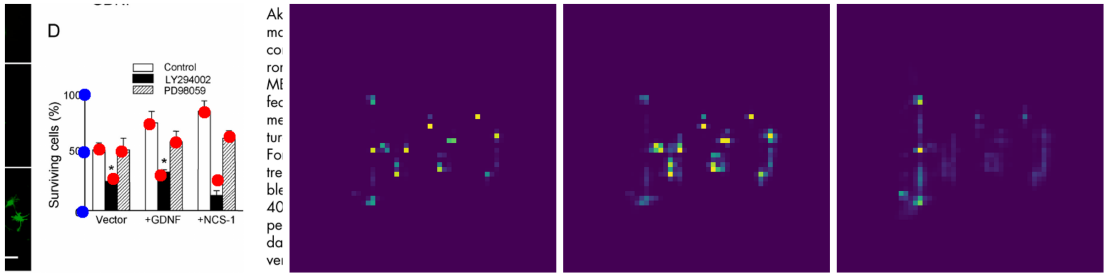
Extracting Values in Charts

Point Proposal Network (PPN): an extensible model for robust data plot value extraction

- Directly predicts location and type of salient data points
- **Extensible to new chart types** (e.g., pie charts, line graphs, etc.) and chart features (e.g., error bars)
- Synthetic data generation enables efficient model pre-training
- **87.05 F1** on *complex* scientific charts





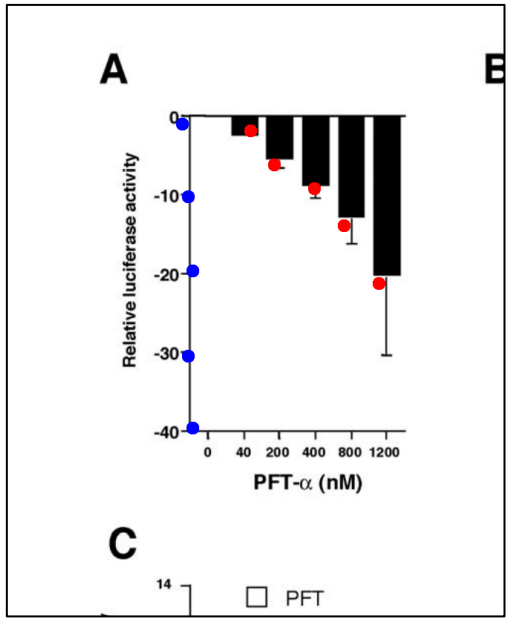


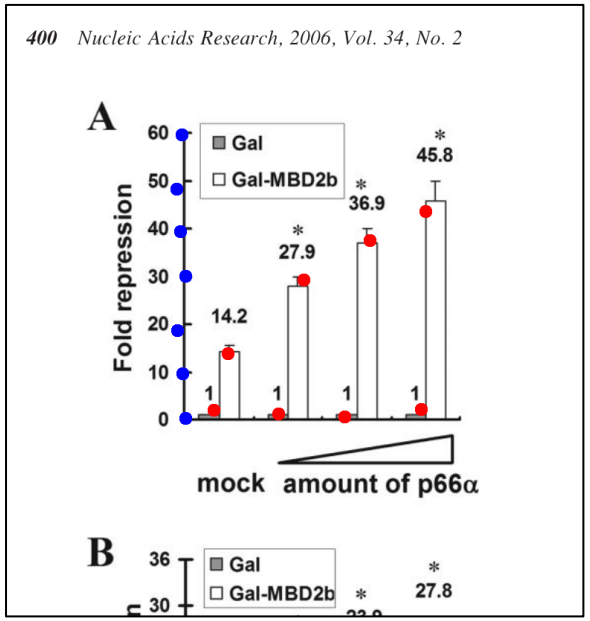
1. Carlos Soto and Shinjae Yoo, “An extensible point-based method for data chart value detection.” arXiv:2308.11788, 2023.

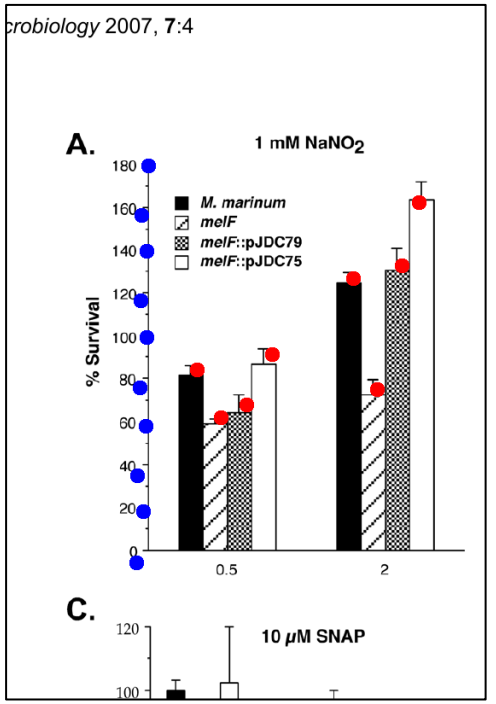
Brookhaven National Laboratory

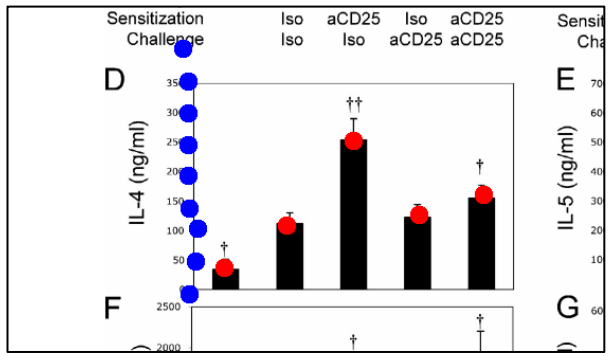
PPN Results and ongoing work

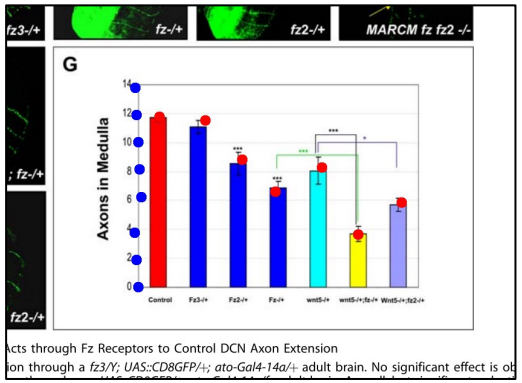
- Working on extensions to new chart elements and chart types
  - e.g., error bars, line graphs, stacked charts
- Building end-to-end tool
  - Integrating with OCR models (with minor modifications)
  - Label alignment and output generation using positional attention

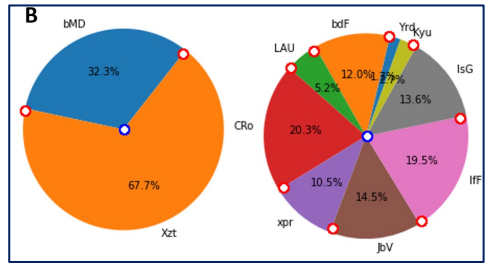








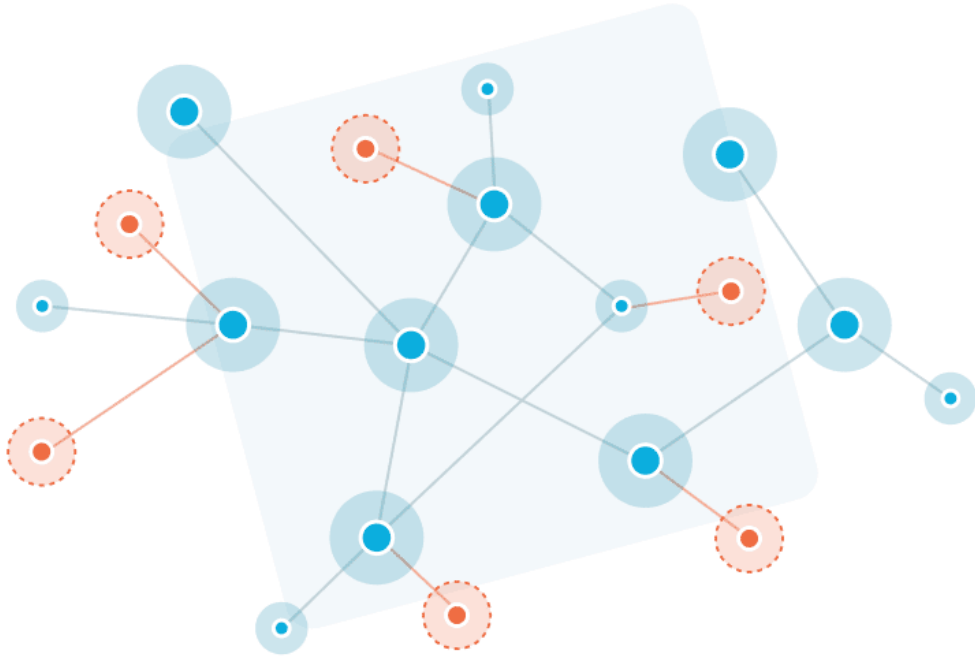




Brookhaven National Laboratory

Vision

- Scalable information retrieval is only beginning
- Build knowledge graphs at scale
  - Explore and interact with NLP
  - Transform and visualize previously disparate data
  - Include programmatic interfaces (APIs) for accessing curated databases, and to feed back into them



csoto@bnl.gov