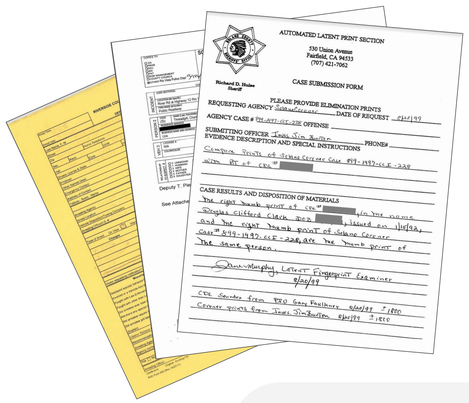
Discovering Structure in Messy Data

Manish Shetty, Aditya G. Parameswaran, Koushik Sen

EPIC DATA lab
UC Berkeley



Messy Data Challenges

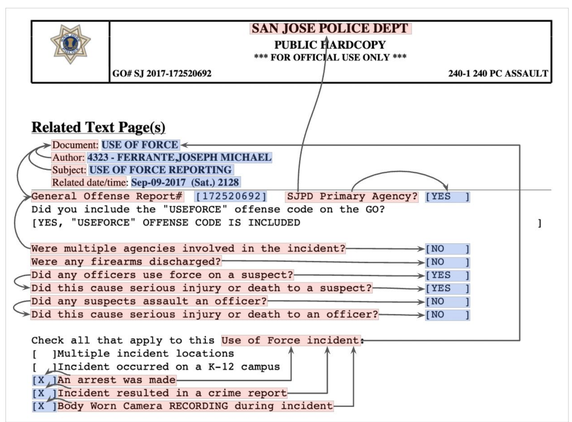
- Duplicates
- Formats
- Unorganized
- Volume
- Unstructured

State-of-the-art

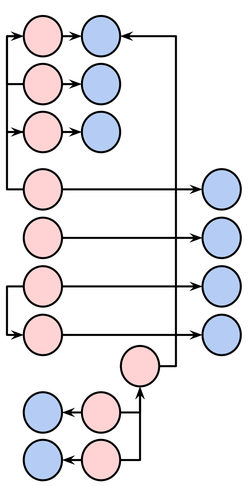
1. Not robust to format changes
2. Do not go beyond identifying isolated entities
3. Do not understand repeated structures
4. Do not completely utilize spatial and semantic cues
5. Require significant manual labeling of data

Most unstructured documents originate from structured data =

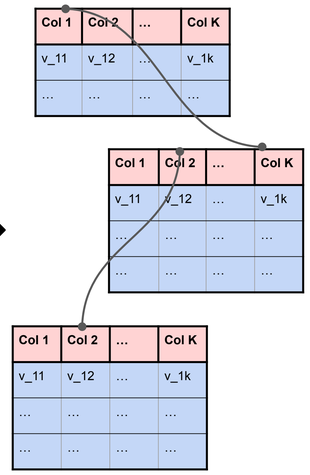
*Can we automatically identify **structure** in **unstructured** data?*



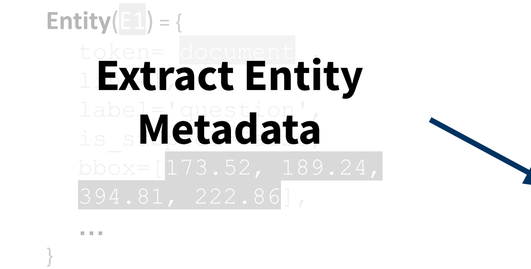
Messy PDF Document



Entity Graph



Structured Tables



Identify Spatial Entity-Entity Relations

Identify Semantic Entity-Entity Relations

Filter "Interesting" Entity-Entity Relations

Domain Specific Language for PDF data extraction

@start Graph<> $\mathcal{G}$:= ConstructGraph($t, \ldots, t$)

Triad ⟨Entity, Entity, string⟩[] t := AllEntityPairs()
| LeftSibling(e)
| RightSibling(e)
| VerticallyAbove(e)
| VerticallyBelow(e)
| *select*

Triad ⟨Entity, Entity, string⟩[] *select* := IsQnA(t)
| IsLabel(lb, lb, t)
| IsMaxDist(d, t)
| IsFixedLoc(x_1, x_2, y_1, y_2, t)
| IsRelatedGPT3(t)
| Conjunction(*select*, *select*)

Entity e
float x_1, x_2, y_1, y_2
string s, lb
int d

@input PDFEntityMatrix

Predictive Synthesis to alleviate annotation overhead!!