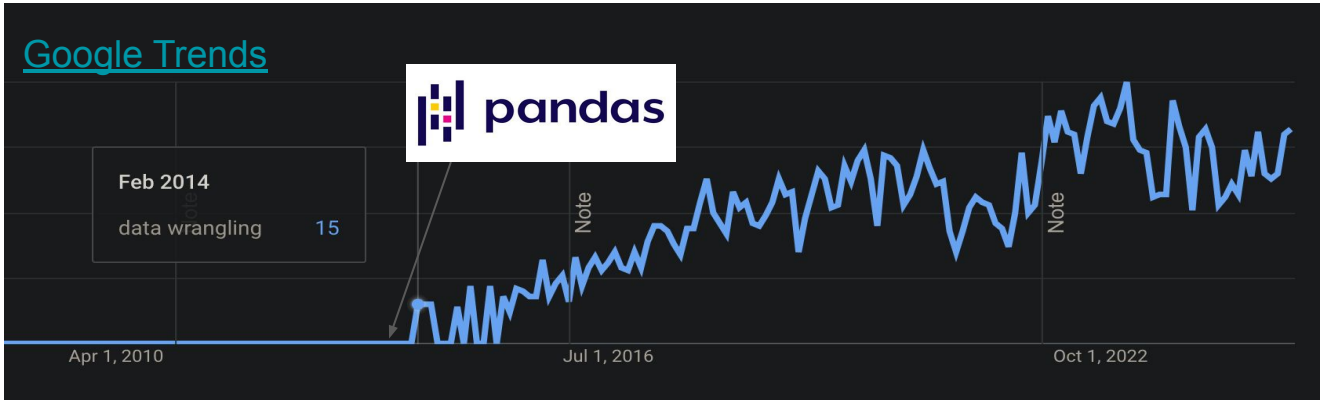
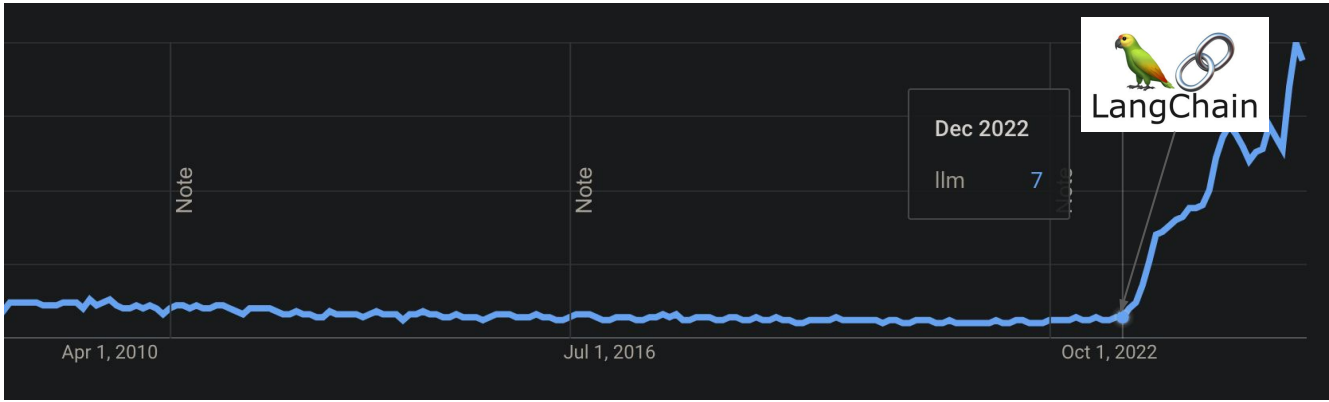
Content Wrangling with LangChain

Collier King
Austin LangChain - AIMUG SXSW
March 2025

Content

~~Data~~ wrangling is the process of converting raw data into a usable form, often for downstream analytics, modeling or presentation.





“Data” vs “Content”

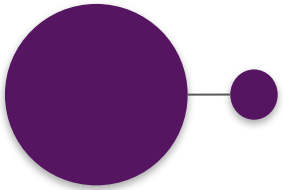
Data - structured, numeric

(ex: relational databases, tabular, spreadsheets, etc...)

Content - unstructured, natural language, audio, video

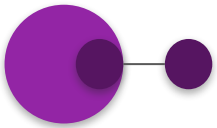
(ex: documents, transcripts, recordings, etc…)

Common Content Wrangling Tasks



Summarization

Condensing lengthy content into shorter summaries and highlights



Extraction

Finding and extracting certain entities and attributes

Supervised

We know what we're looking for and we're telling the LLM to go get it

Unsupervised

We aren't sure what we're looking for and want to see what the LLM brings to us

Classification

Given a choice of categories, determine which one(s) a piece of content falls under

Caveats & Challenges

Content Wrangling is not one size fits all

For best results Prompts should be fine-tuned for domain areas

Checks should be added for potential Hallucinations

Notebook Demo

https://github.com/CollierKing/langchain-content-wrangling