

The background of the slide is a composite image of Earth and the Moon in space. The Earth is shown as a large, curved horizon on the left side, with blue oceans and white clouds. The Moon is visible in the upper right corner, showing its craters and grey surface. The sky is a deep blue with scattered white stars.

# TAMING THE WORLD WIDE WEB

Challenges faced when dealing with 200k+ websites

Lexis Nexis

Eric Platow – Senior Director of Data Science

# TOPICS

---

Project Introduction

Extraction Engine

Guidelines

Tools

High Level Process

Summary



# INTRODUCTION



- **Problem Definition:**
  - ▢ Needed to find and update ~1M attorney biographical records

# DEADLINE

---






# WHY?

---

Who would you like to analyze using language analytics?

*Enter the name of an attorney*

Attorney 





# HELP!

---

- This talk will walk you through the journey from concept to iteration to designing your next large scale data collection project



# GUIDELINES

---

Must Be Repeatable

Check Data Monthly

Must Be Automated

Easy To Start Processing

Must Scalable

Handle Hundreds of Thousands of Websites

Must Be Extendable

To Other Countries and Company Types

Must Be Reusable

For Unforeseen General Scraping Requests



# THE WORLD WIDE WEB SPANS...





# RABBIT HOLES

---





# DEFINE RULES



## Legal Verification:

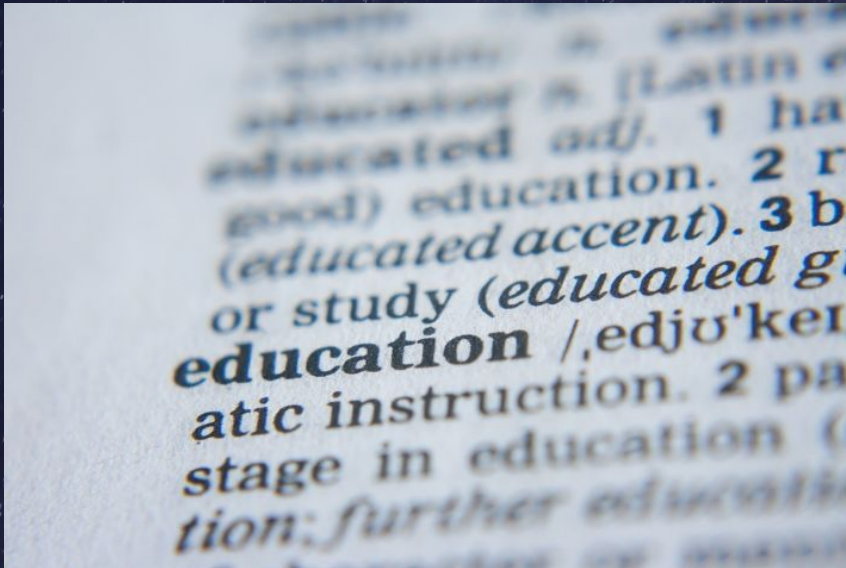
- Terms And Conditions
- PII
- Copyrighted Information

## Being A Good Citizen:

- Scrape Times
- Robots.txt
- Max Time On Site
- Exclusion Paths
- Strict Paths



# DEFINE RULES



Must Define What To Do:

- Redirected Sites
- Squatted Sites
- Domain For Sale
- Inappropriate Sites
- Subdomains



# DEFINE RULES



## In/Out of Scope:

- Captcha
- Antibot
- Logins
- Requests / Headless / Full Browser



# HIGH LEVEL PROCESS





# AUTOMATED DATA EXTRACTION

---

Classify the site

Start simple

Core terms or even pre-classified

Classify the page ... is it a page type you care about or not

Start simple

Define Data Priorities

Contact information

Education

Multiple Extraction Layers



# EXTRACTION LAYER – HTML



## Kitty Kat

Partner

Washington, DC

(212)-555-1212

### Education:

- Fishfood University – 1998
- Catnip University – 2000 (JD)

### Admissions:

- Idaho 2001

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

# EXTRACTION LAYER 2 – NLP

Named Entity Recognition and Natural Language Processing Layer

NER – Named Entity Recognition

Spacy: <https://spacy.io/>

Pyap: <https://pyap.io/>

FYI: <BR>

NLP – Natural Language Processing

Select the Co

Bert/Ro

Que

What

Layer NER with

Annotating Exam

Prodigy: [Train Models](https://prodi</a></p></div><div data-bbox=)

Compare and Combine Extracted Information

Cost and Speed Considerations





# EXTRACTION LLM

---

Security

Supply Content and Ask Questions

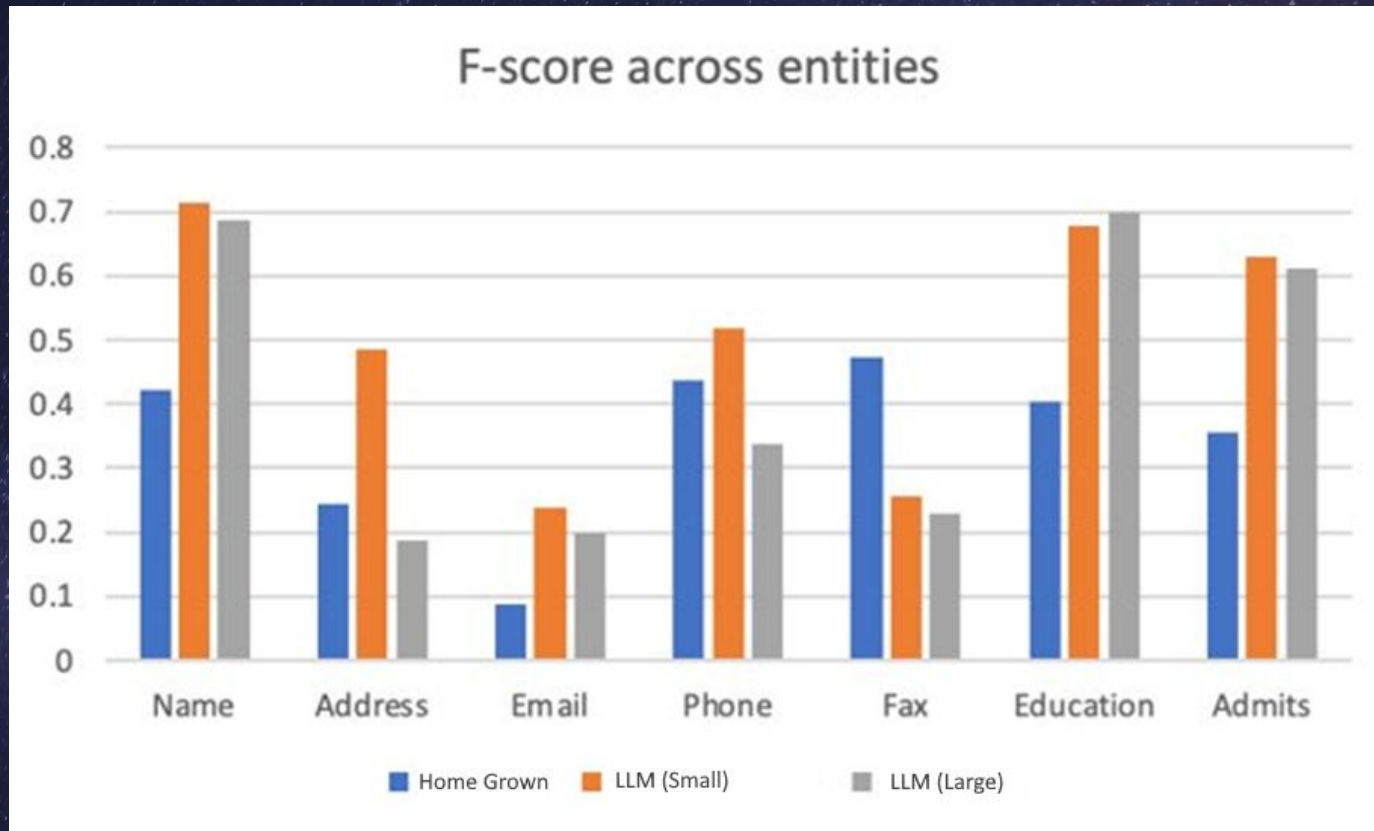
Prompt Engineering

Not a FULL solution either!



# EXTRACTION LLM

---





# AUTOMATED DATA PROCESSING

---

How to process the data

Wild Wild West

Data Cleanliness is non-existent

Fuzzy matching to known lists where possible

```
textdistance.sorensen_dice.normalized_similarity
```

```
textdistance.lcsstr.normalized_similarity
```

```
fuzz.token_set_ratio
```

Fuzzy matching to determine if we had the information

Decision Logic Tree

Defined outcome groups

Document Document Document!

Report What We Did



# PROJECT OUTCOME

---

- **\$3.7M** of Editorial Avoidance
  - Avoided hiring / training of **~400** people
  - **COMPLETELY** changed the way our editorial team works
- They now work on exceptions or the higher value investigation of issues



# TOOLING

---

- URL Cleanup:
  - Cleanliness of source URLs
  - Htp or <http:///> or <http://email@domain.com> ....
- Site Checker:
  - Validate the site exists, url redirections, validity of site

URL	actual_url	site_exists	is_intl_domain	site_redirected	site_valid	site_valid_terms	website_found	website_title
tobolowskylaw.c	<a href="http://tobolowskylaw.com">http://tobolowskylaw.com</a>	TRUE	FALSE	FALSE	TRUE	law,lawyer,legal	TRUE	TOBOLOWSKY
collinfamilylaw.c	<a href="http://collinfamilylaw.com">http://collinfamilylaw.com</a>	TRUE	FALSE	FALSE	TRUE	attorney,law,lawyer	TRUE	Collin County

- Stats about Completion And Automated Ticketing
  - # Pages found
  - # Documents downloaded
  - Site still valid
- Sites Change; Get Blocked
  - What you get today you may not get tomorrow
  - Alarms if num pages drop



# TOOLING

Researching URLs is becoming MORE  
and MORE dangerous

URL link validation tools

[URLScan.io](https://urlscan.io)

Air Gapped Virtual Machines





# SUMMARY

---

## Remember

- Web Started in 90s
- Sophisticated Sites
- Define Rules
- Asynchronous Architecture
- Define Data Extraction Techniques
- Start Simple
- Assume Sites Will Change + Stop Working

