Jet Propulsion Laboratory
California Institute of Technology

Transforming DARPA and NASA JPL’s Deep Web Research into Competitive Intelligence

Kyle Hundman, Data Scientist, Office of the Chief Information Officer

Agenda

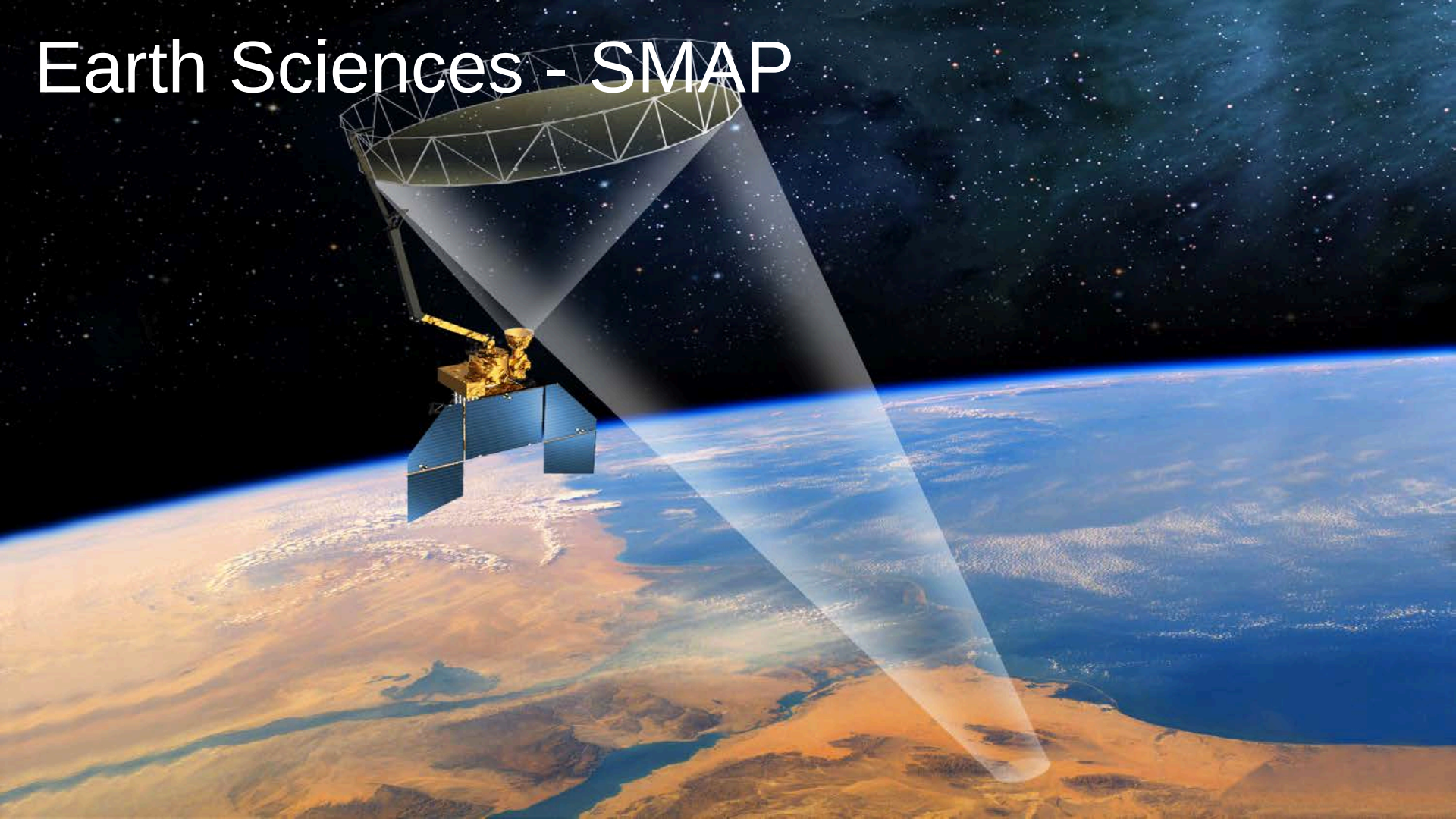
1. **Background**
2. Who cares?
3. Under the hood
4. How it translates
5. Challenges

JPL

- Pasadena, CA
- 1 of 10 NASA centers
- Operated by Caltech
- ~5,500 employees
- 3 domains of expertise

Robotic Spacecraft - Curiosity

Earth Sciences - SMAP



Astronomy - Keplar

Memex – DARPA's search engine for the Dark Web

America needed a way to make sense of gargantuan icebergs of raw and seemingly unconnected data.

- **Deep web (95%)**
- **Human trafficking, weapons**
- **Domain monitoring**

Agenda

1. Background
2. **Who cares?**
3. Under the hood
4. How it translates
5. Challenges

Memex Impact

DARPA Research: Trafficking industry spent **$250M** to post over **60M** ads over two-year period

Memex is in use in several law enforcement agencies and has led to 100+ arrests

NYDA: Memex has contributed to 20+ active investigations and 9 open indictments

http://humantraffickingcenter.org/posts-by-htc-associates/memex-helps-find-human-trafficking-cases-online/

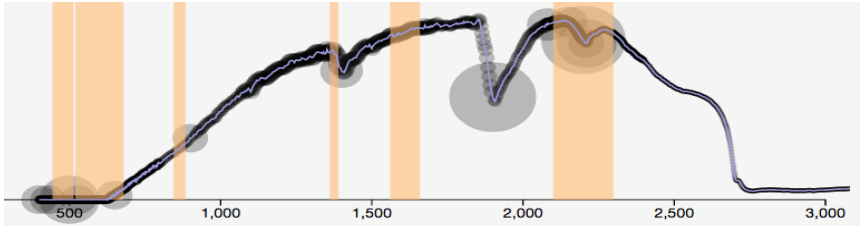
Panama Papers

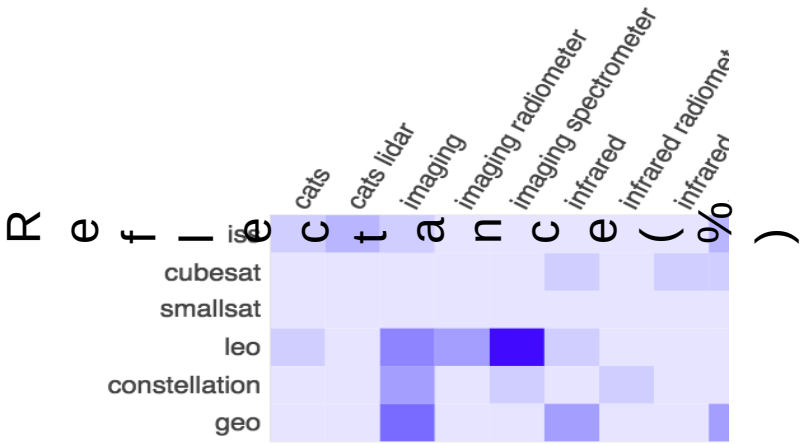
- 2.6 terabytes
- 11m documents
- >100 publications

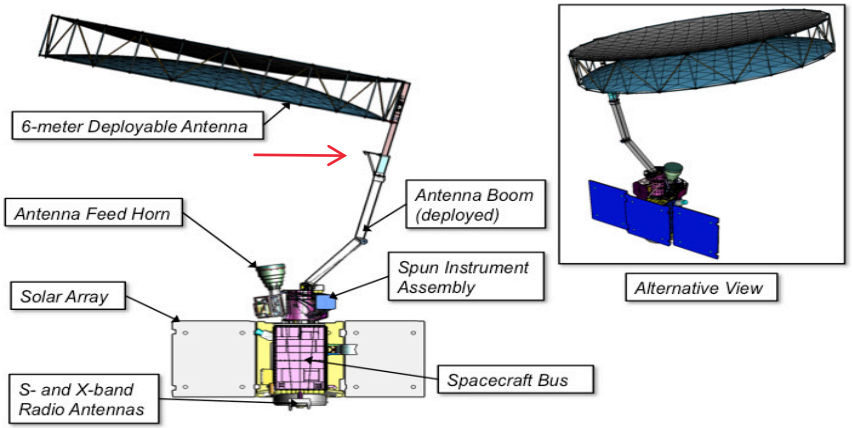
“All the software used was open source, tweaked to suit the reporters’ needs. The search tool, allowing reporters to hunt for names like Putin or places like the British Virgin Isles, was based on **Apache Solr**, used by a large number of search-heavy organizations, including DuckDuckGo, a privacy-focused tool. **Solr** was combined with **Apache’s Tika**, an indexing software that can parse different file types, be they PDFs or emails as in the Panama Papers, drawing out the text from the non-essential data.”

http://www.forbes.com/sites/thomasbrewster/2016/04/05/panama-papers-amazon-encryption-epic-leak/?utm_campaign=ForbesTech&utm_source=TWITTER&utm_medium=social&utm_channel=Technology&linkId=23087770#7dfd3a691df5

- Scientific Literature
- Proposals
- Solicitations



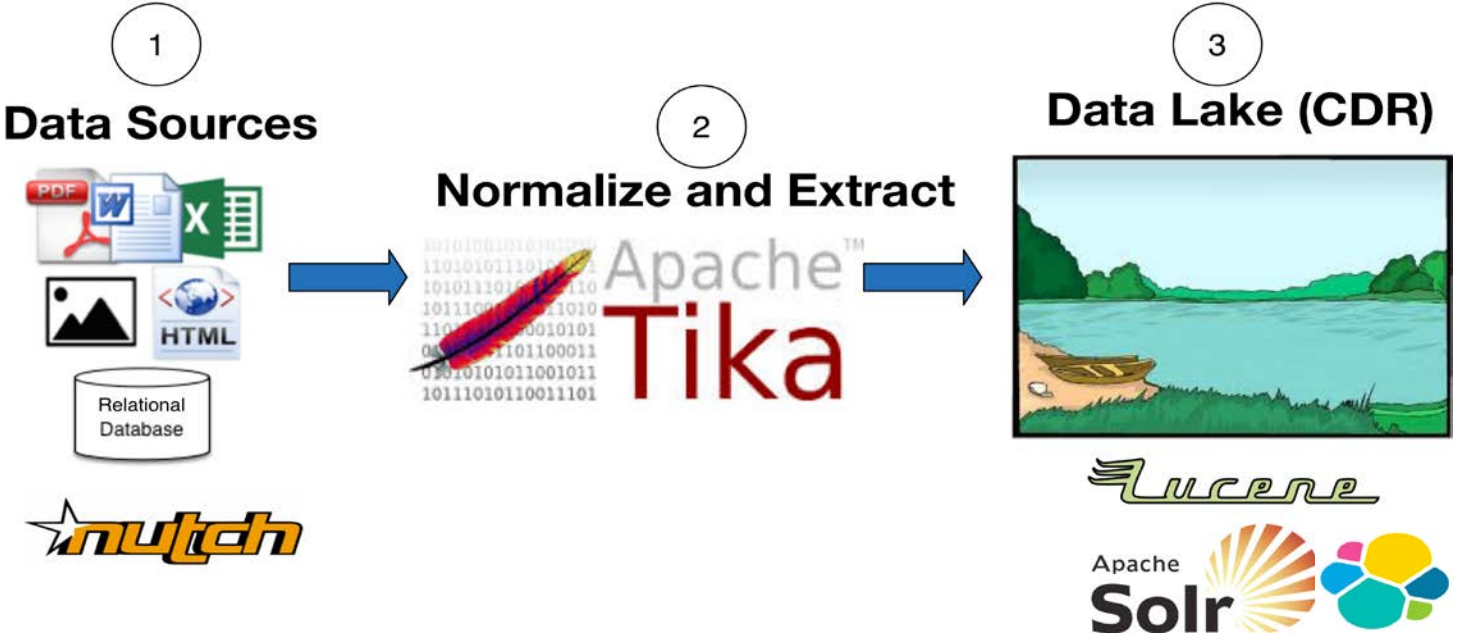




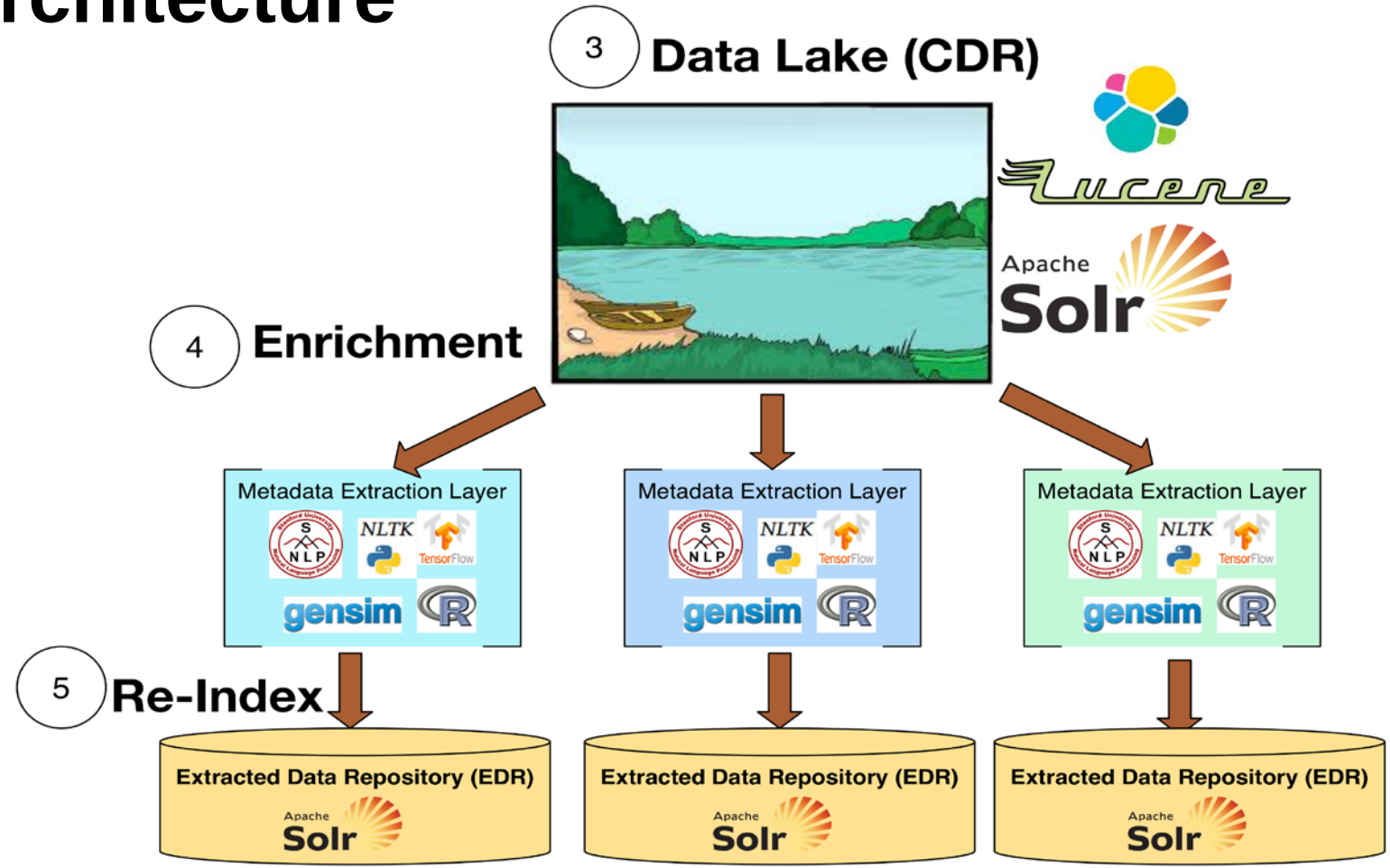
Agenda

1. Background
2. Who cares?
3. **Under the hood**
4. How it translates
5. Challenges

Architecture



Architecture



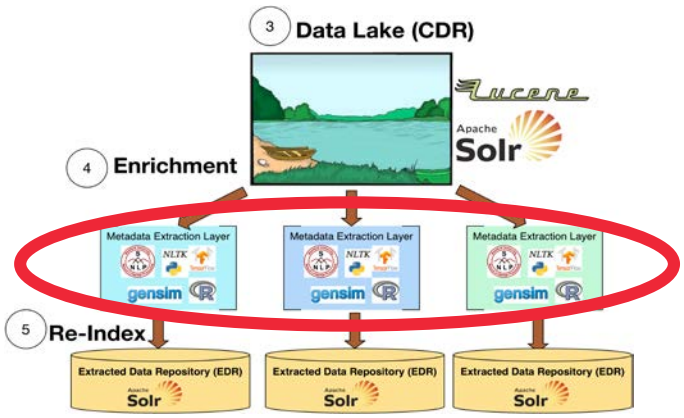
Example from DARPA Memex

Question:

Is this ad associated with human trafficking?

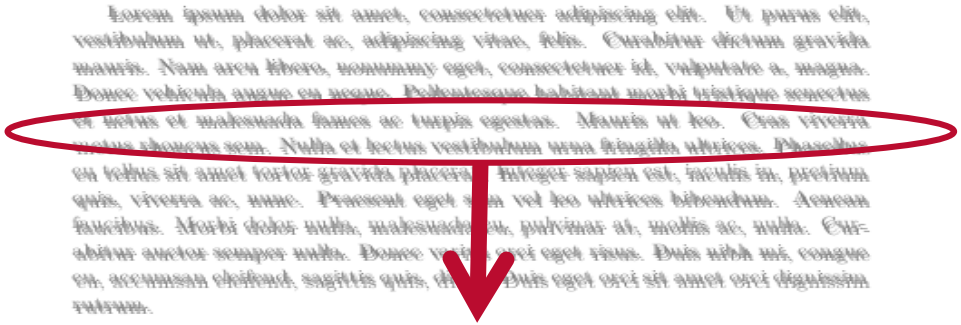
Goal:

Extract indicators that suggest an ad is related to human trafficking



- Traffickers move victims from place to place
- Underage victims
- Traffickers have multiple victims simultaneously

Example – Stanford DeepDive



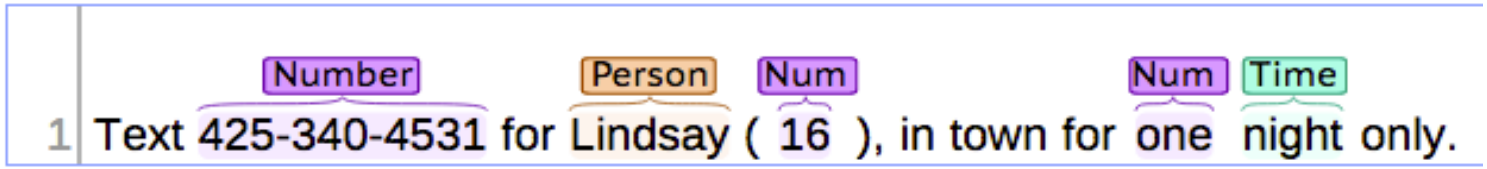
"Text 425-340-4531 for Lindsay (16), in town for one night only."

Named Entity Recognition:

1 Text 425-340-4531 (Number) for Lindsay (Person) (16 (Num)), in town for one (Num) night (Time) only.

Example (cont.)

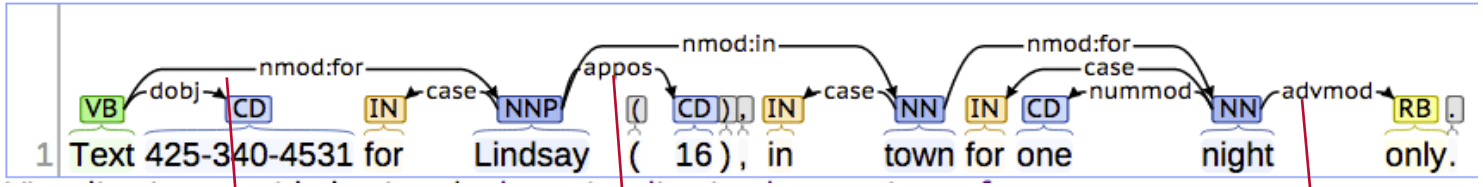
Named Entity Recognition:



Potentially…

- Trafficker ID
- Victim
- Age
- Timeframe

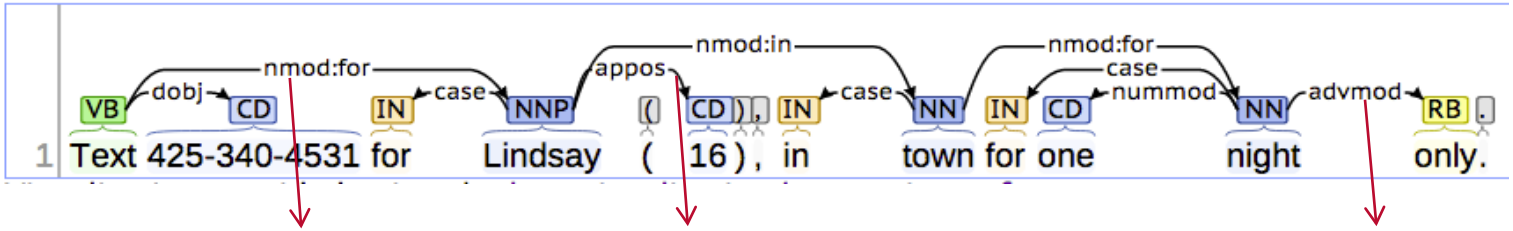
Enhanced Dependencies:



- Trafficker's victim
- Victim's age
- One is referring to night

Example (cont.)

Enhanced Dependencies:



DeepDive Guesses:

Lindsay is a victim of 425-340-4531

Lindsay is 16

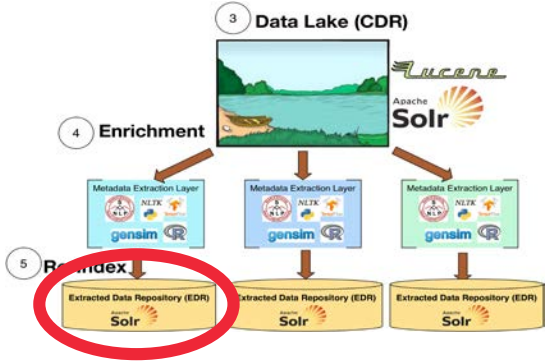
This is an indication of travel

Human validation:

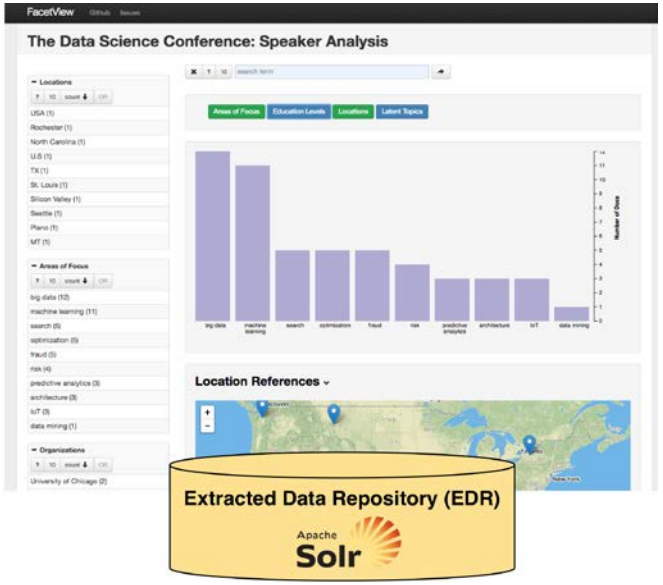
True **True** **True**

Automatic Model Refinement

Interfaces



6 Apps



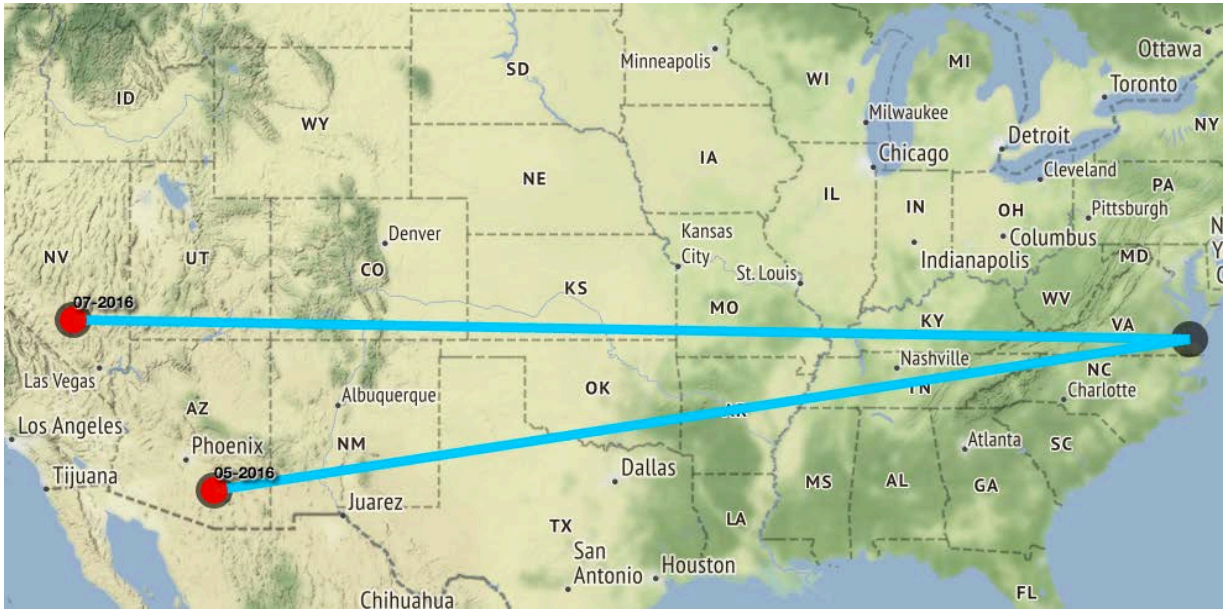
Given _____ data, what relationships, trends, similarities, and outliers exist?

- Expediting and automating data pipelines & analysis

Interface Example

Phone numbers + locations + time

- Separate trafficking from prostitution
- Predict movements



Agenda

1. Background
2. Who cares?
3. Under the hood
4. **How it translates**
5. Challenges

Problems to Address

1. Size + Lack of Structure + Inaccessibility of data

internal + external
text, websites, documents, and images
Deep, Dark data, archaic systems

2. Thorough manual analysis is infeasible

Resulting in missed opportunities to identify **relationships, trends, gaps, and outliers** in internal and external data.

Leveraging at JPL

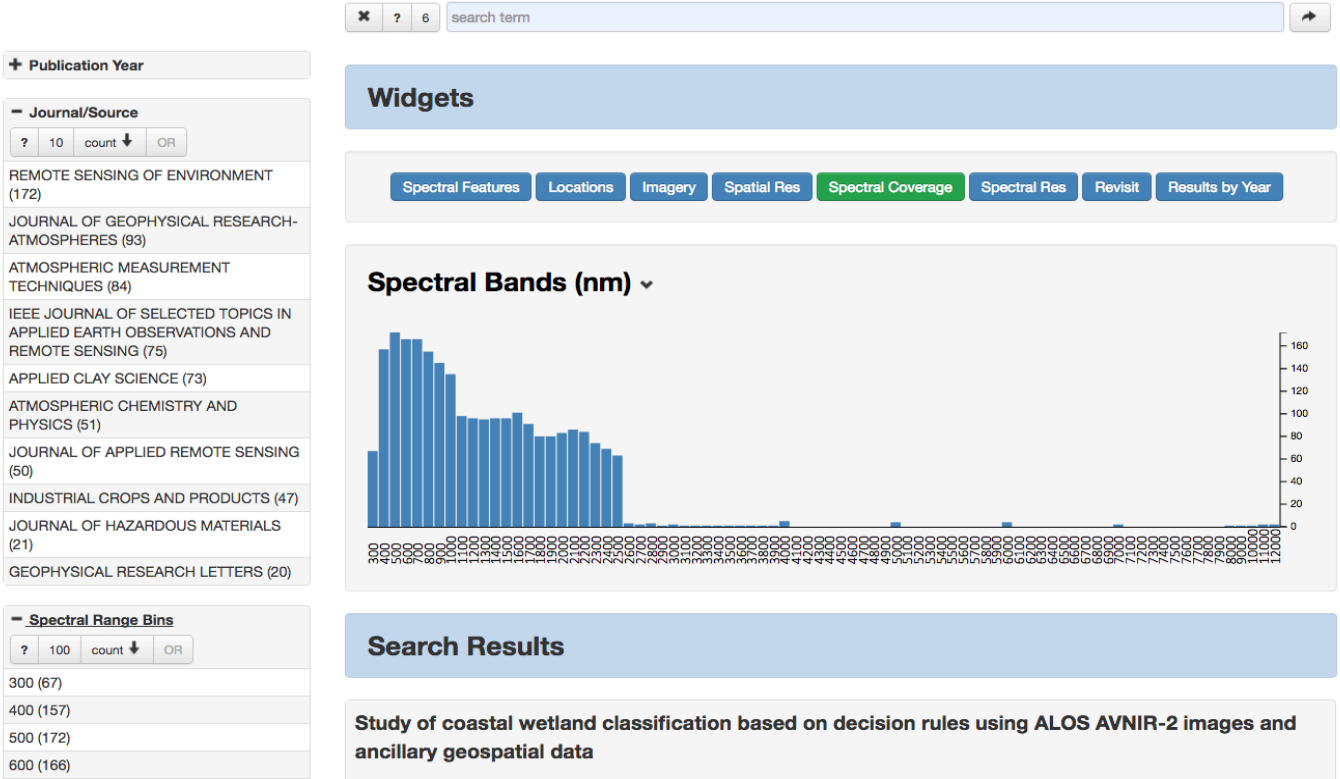
- Competitive intelligence in **Remote Sensing**

- Massive amounts of unstructured data
  - Earth Sciences has **485 research tasks** with **1,000+ authored papers** in 2016 so far
  - 12m docs at JPL

- Extraction relevant information such as:
  - Measurements (e.g. spectral coverage, spatial resolution)
  - Satellite mentions
  - Surveying methods

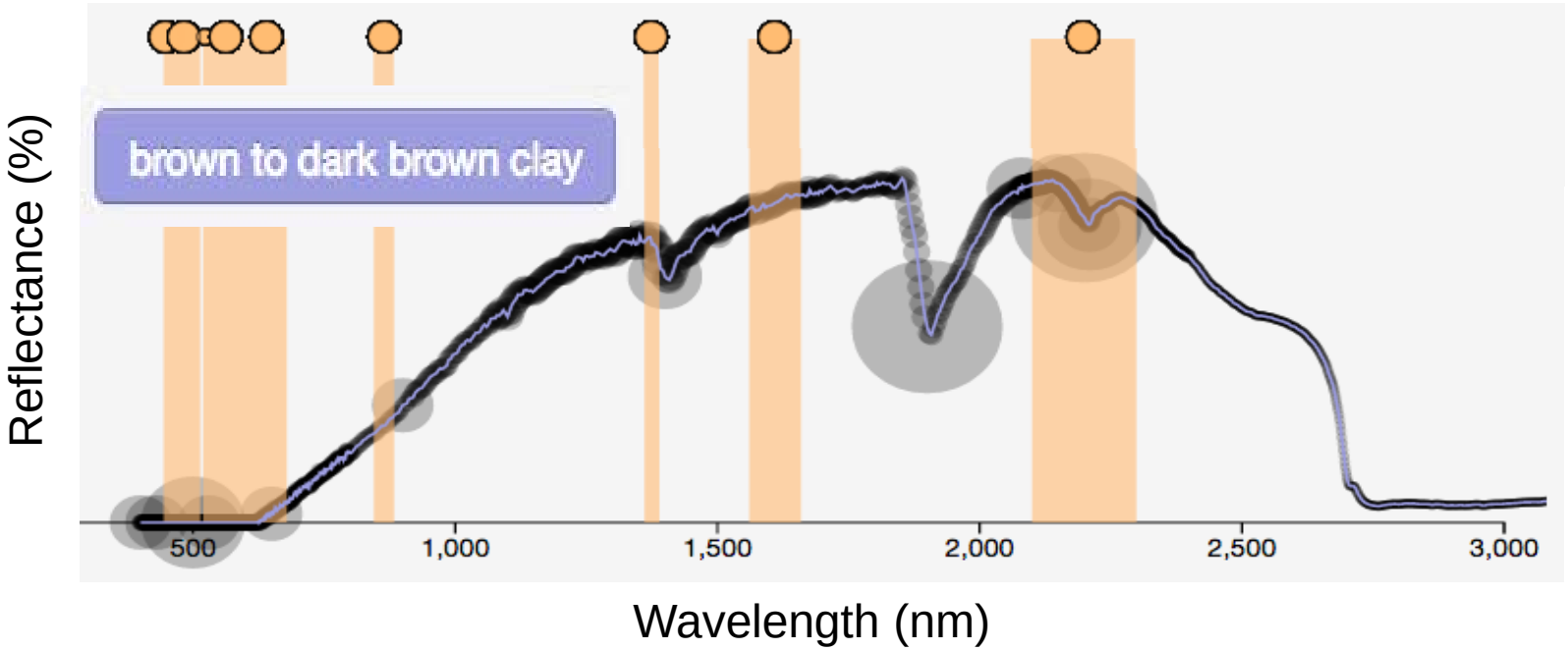
Example in Remote Sensing

- 2,500 Publications from Remote Sensing Domain
- Trends, outliers through faceted search



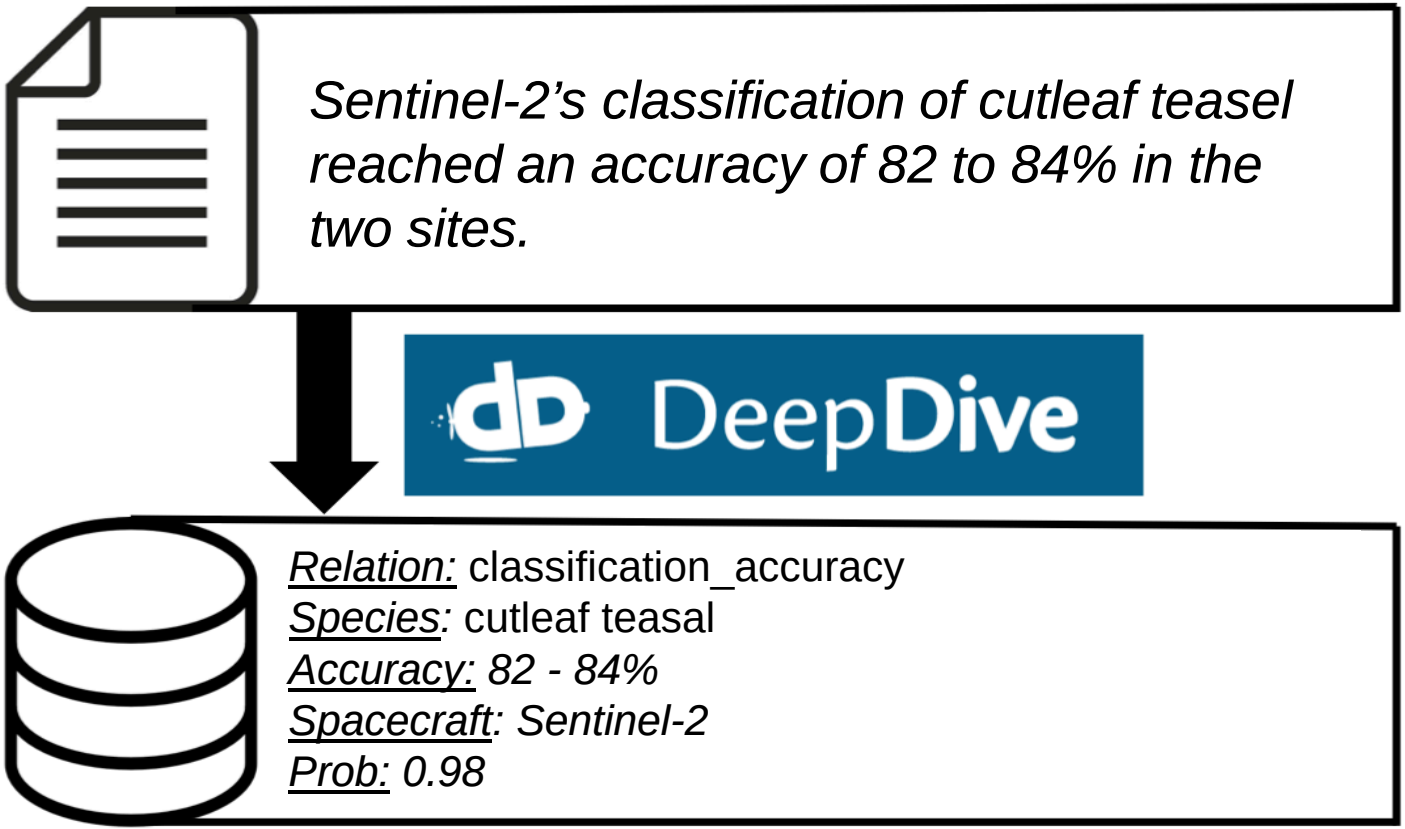
Example in Remote Sensing (cont.)

- identify measurement gaps and potential science opportunities



Remote Sensing

- Curate relevant information



Agenda

1. Background
2. Who cares?
3. Under the hood
4. How it translates
5. **Challenges**

Uncertainty of Results

How do I know these results are accurate?

- You don’t
- Sanity checks
- Quantify when possible
- Understand research results
- Scale and diversity help
- Pay for labels

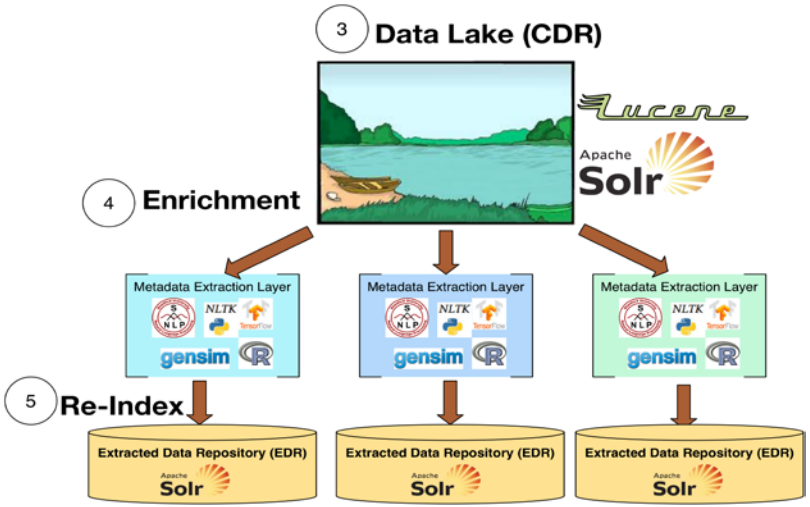
Expertise and Cost

- Diversity of skills needed across pipeline
- Some degree of customization is required to gain value
- Open source generally requires more development chops
- Need subject matter expertise to guide development
- Rapidly evolving field

Data Engineer

Data Scientist

Web Developer



Results

1. Automated, large-scale processing of unstructured data (10s of millions of ads for Memex)

1. Move effort away from data and towards analysis and hypothesis testing

1. Enabling non-technical users to broad or deep analysis

=

Insightful & Informed Decision-Making

Links / Acknowledgments

JPL Office of the Chief Information Officer (OCIO)

Topic/Tool	Link
Facetview	https://github.com/chrismattmann/facetview
JPL Memex Code Base	https://github.com/memex-explorer
Apache Nutch	http://nutch.apache.org/
Apache Tika	https://tika.apache.org/
Deep Dive	http://deepdive.stanford.edu/
CoreNLP Demo	http://nlp.stanford.edu:8080/corenlp/process