Finding relations in documents with IEPY

machinalis

Daniel F. Moisset – dmoisset@machinalis.com

tw: @dmoisset

This is a talk about...

A (very brief) intro about IE

Example applications

IEPY, an open source tool to do it in python

`https://github.com/machinalis/iepy`

About me

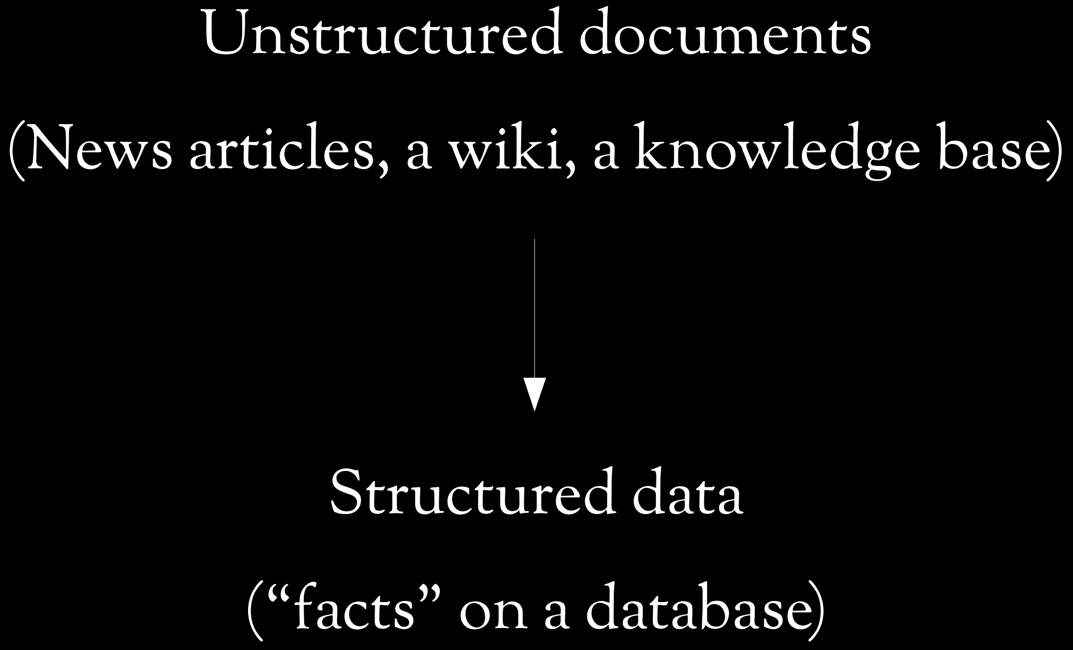
I am a Computer Scientist

I am a cofounder and technical leader at Machinalis

Machinalis (http://www.machinalis.com) provides development services for Machine Learning applications.

I participated in the development of IEPY

Information Extraction



Unstructured documents

(News articles, a wiki, a knowledge base)

Structured data

("facts" on a database)

Some applications

Given…	**Obtain a table of**
A collection of medical papers	Genes and proteins expressed by them
Wikipedia	Relevant people and their dates of birth
A set of business articles	Company acquisitions and fundings
Social media posts	Foodborne disease outbreaks

In general…

Extract *RELATIONS*

(tuples of a given arity expressing some known concept/meaning)

Between *ENTITIES*

(People, places, organizations, dates, amounts of money, …)

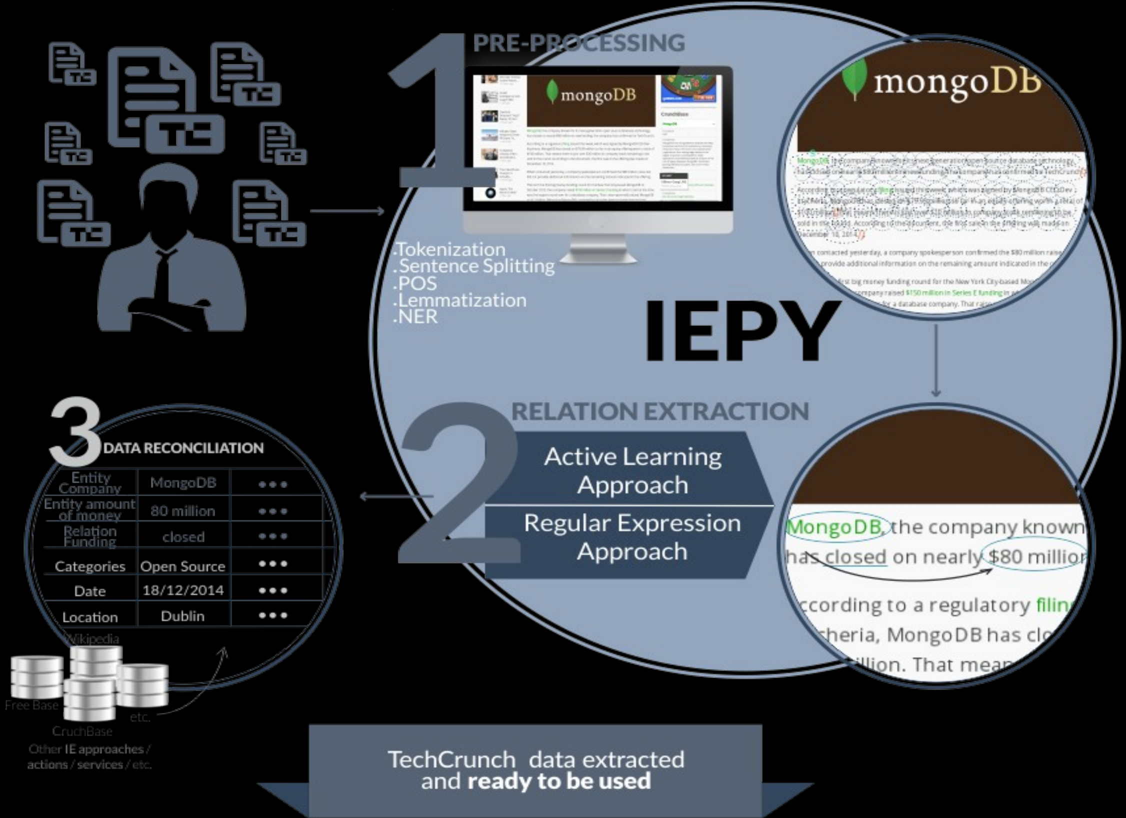
From *DOCUMENTS*

(text and maybe some associated metadata. Also possible over other media, not discussed in this talk)

Example: **Pandas** is a software library written for the **Python** programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. **Pandas** is free software released under the three-clause **BSD license**.

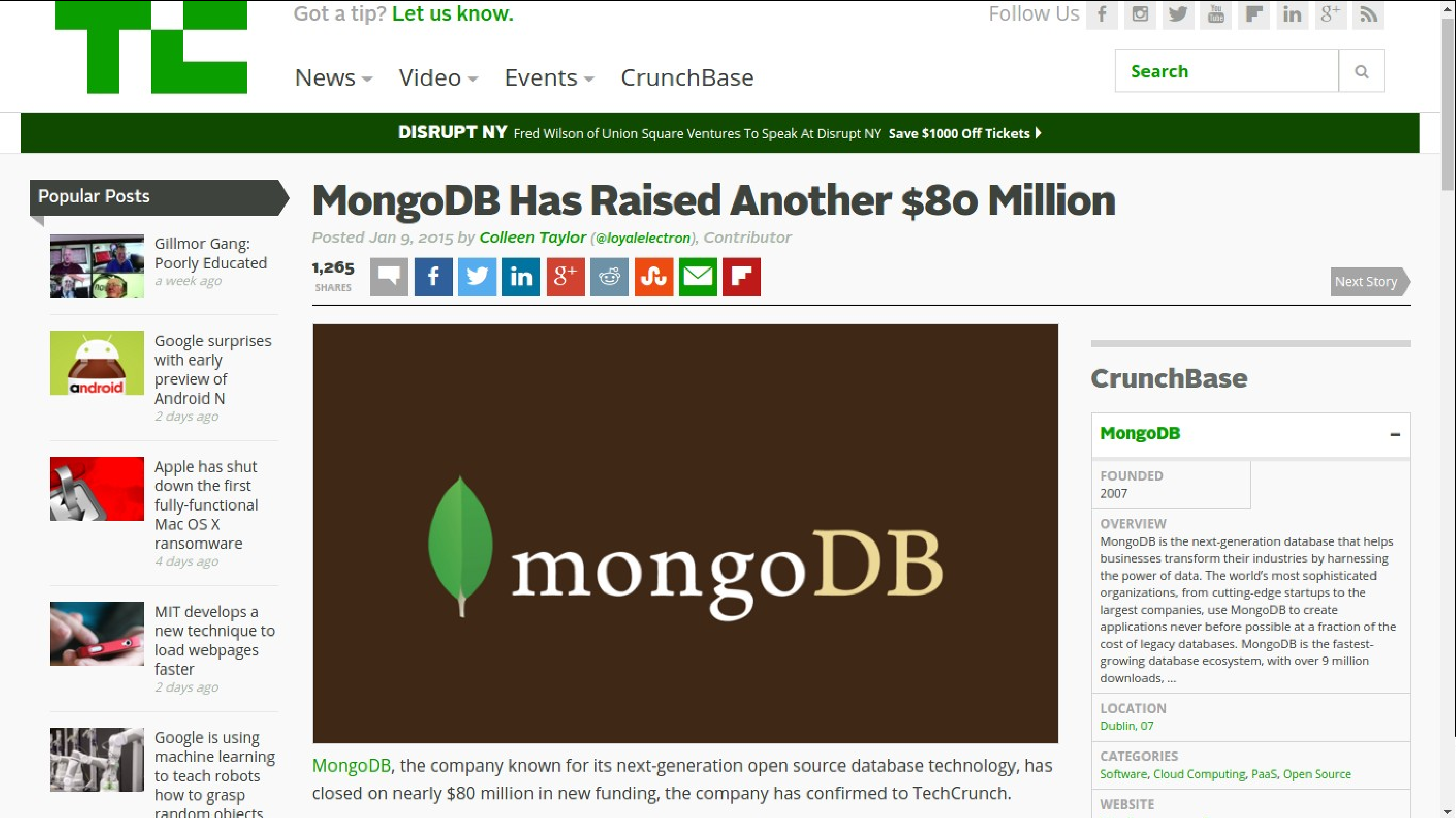
An example: iepycrunch

http://iepycrunch.machinalis.com/



Entity Company	MongoDB	•••
Entity amount of money	80 million	•••
Relation Funding	closed	•••
Categories	Open Source	•••
Date	18/12/2014	•••
Location	Dublin	•••

TechCrunch data extracted and **ready to be used**



Popular Posts

Gillmor Gang: Poorly Educated
a week ago

Google surprises with early preview of Android N
2 days ago

Apple has shut down the first fully-functional Mac OS X ransomware
4 days ago

MIT develops a new technique to load webpages faster
2 days ago

Google is using machine learning to teach robots how to grasp random objects

MongoDB Has Raised Another $80 Million

Posted Jan 9, 2015 by Colleen Taylor (@loyalelectron), *Contributor*

1,265 SHARES

Next Story

MongoDB, the company known for its next-generation open source database technology, has closed on nearly $80 million in new funding, the company has confirmed to TechCrunch.

CrunchBase

MongoDB

FOUNDED
2007

OVERVIEW
MongoDB is the next-generation database that helps businesses transform their industries by harnessing the power of data. The world's most sophisticated organizations, from cutting-edge startups to the largest companies, use MongoDB to create applications never before possible at a fraction of the cost of legacy databases. MongoDB is the fastest-growing database ecosystem, with over 9 million downloads, ...

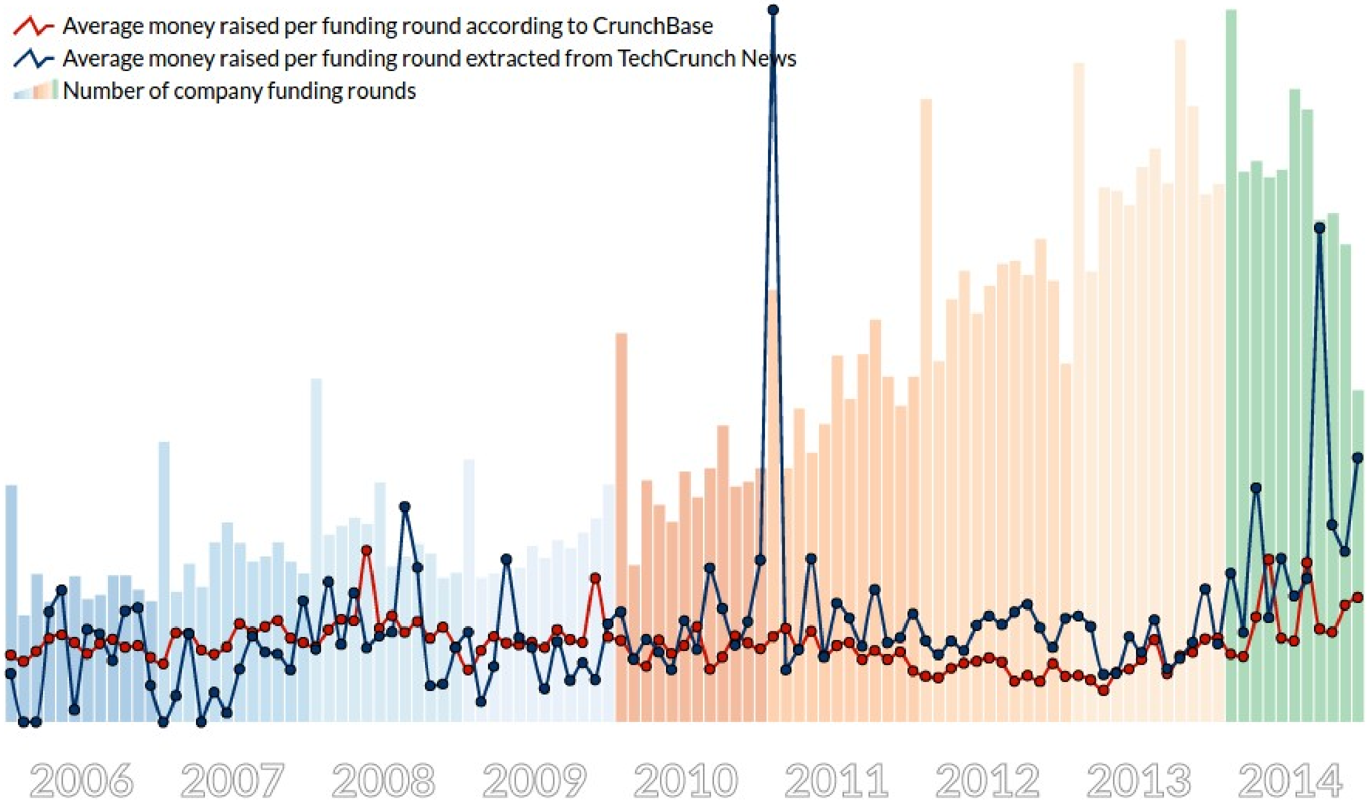
LOCATION
Dublin, 07

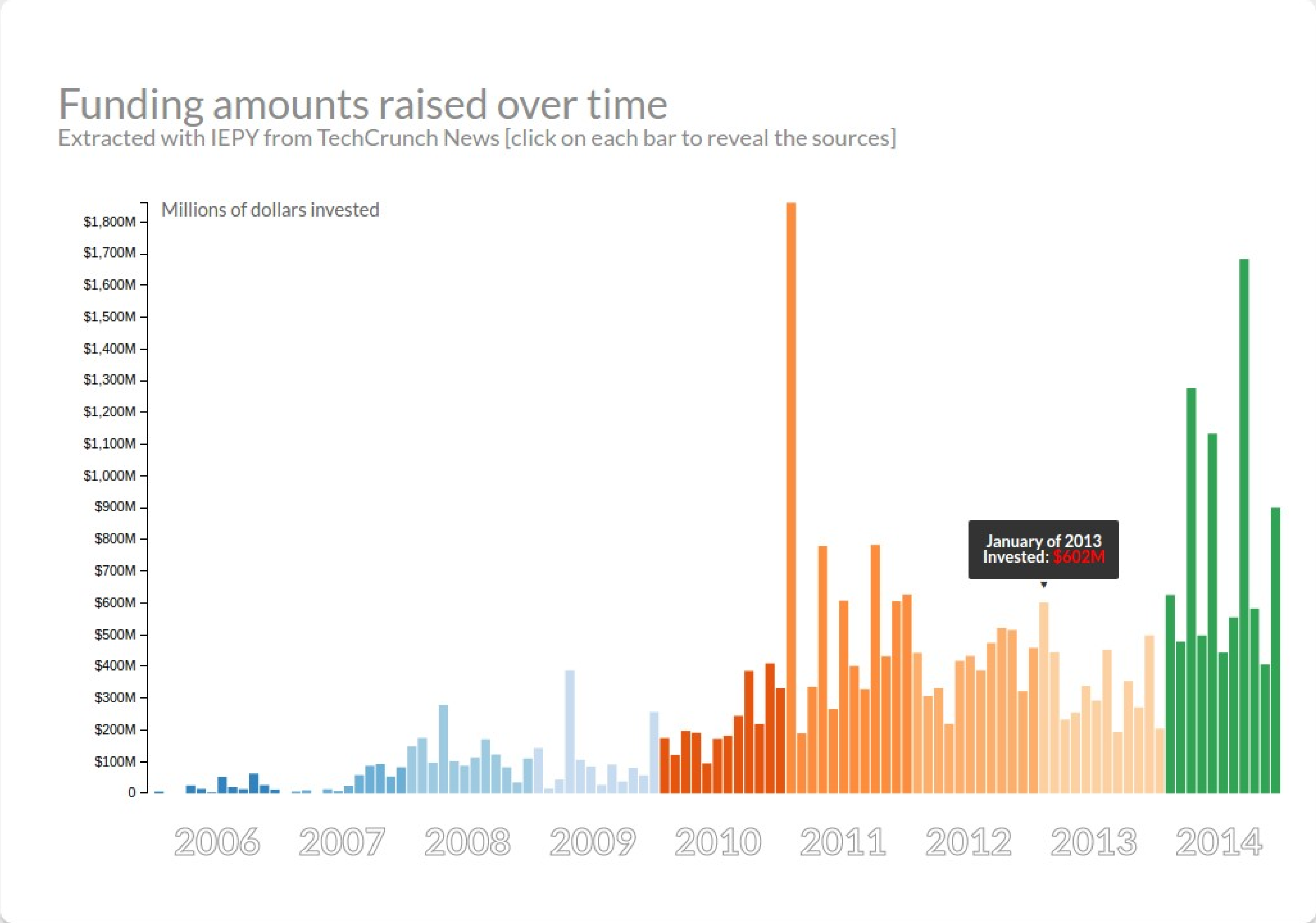
CATEGORIES
Software, Cloud Computing, PaaS, Open Source

WEBSITE

Funded Companies vs Average money raised

Both the VC Industry and the specialized press are discussing trends in funding rounds in recent years.
Here there are some differences.

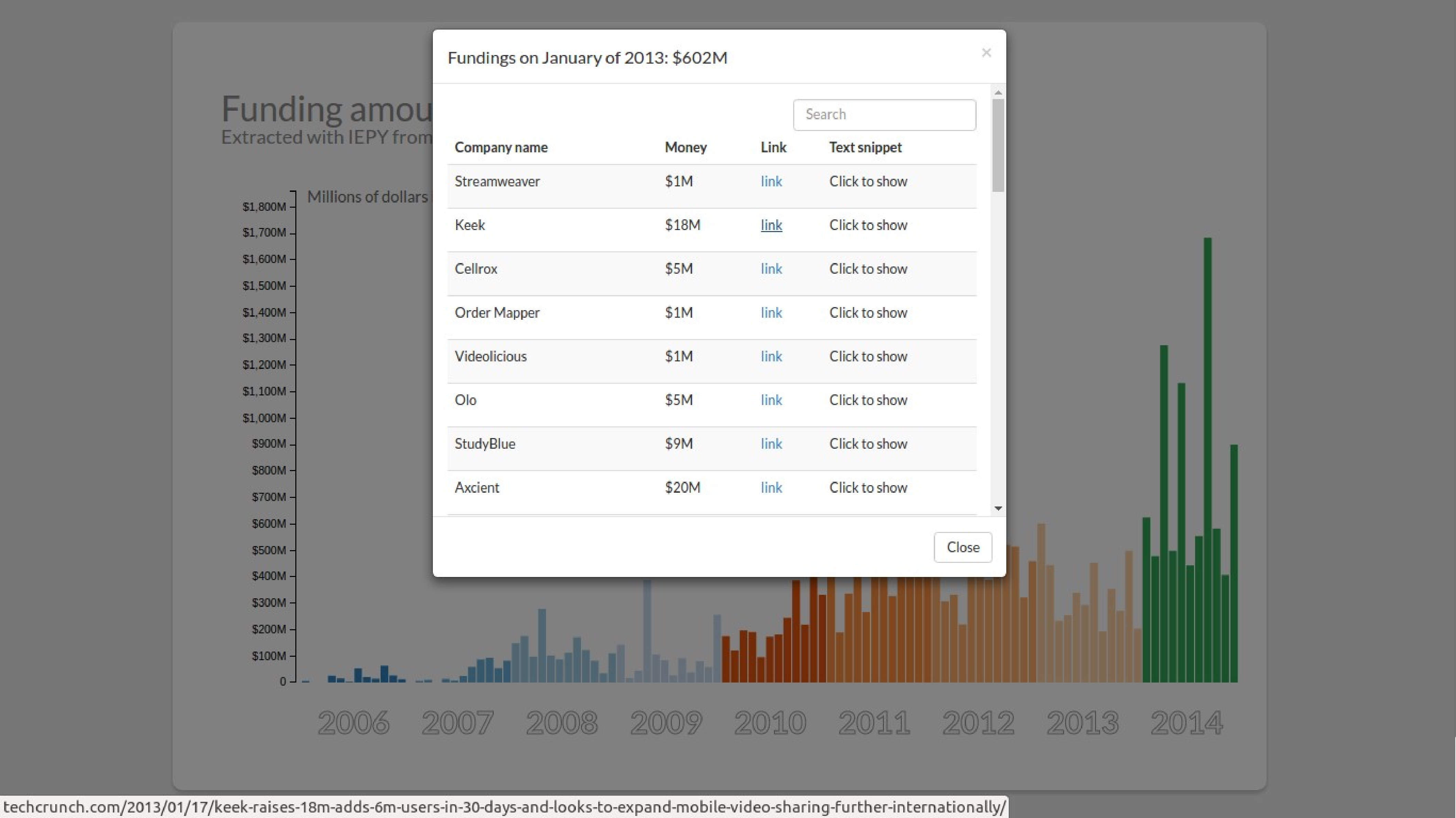




Funding amounts raised over time

Extracted with IEPY from TechCrunch News [click on each bar to reveal the sources]

January of 2013
Invested: $602M



Fundings on January of 2013: $602M

Company name	Money	Link	Text snippet
Streamweaver	$1M	link	Click to show
Keek	$18M	link	Click to show
Cellrox	$5M	link	Click to show
Order Mapper	$1M	link	Click to show
Videolicious	$1M	link	Click to show
Olo	$5M	link	Click to show
StudyBlue	$9M	link	Click to show
Axcient	$20M	link	Click to show

IEPY

https://github.com/machinalis/iepy

Python based tool + framework for creating IE applications

- Some parts of the pipeline have defaults that can be replaced
- Supports automatic, manual, or hybrid approaches
- Includes a web UI for manual interaction

Python (+ external components), BSD open-source

IEPY pipeline: Load the documents

https://github.com/machinalis/iepy

IEPY manages its own database for storing documents

- A simple CSV importer is included
- You can write others or convert to CSV
- It is possible to do incremental updates

Some work is typically required here given the variety of possible inputs

IEPY pipeline: Preprocessing (1)

- Tokenization
- Segmentation in sentences
- Lemmatization
- Part of Speech tagging

IEPY uses off-the-shelf tools like Stanford NLP toolkit, you can add/change steps.

IEPY pipeline: Preprocessing (2)

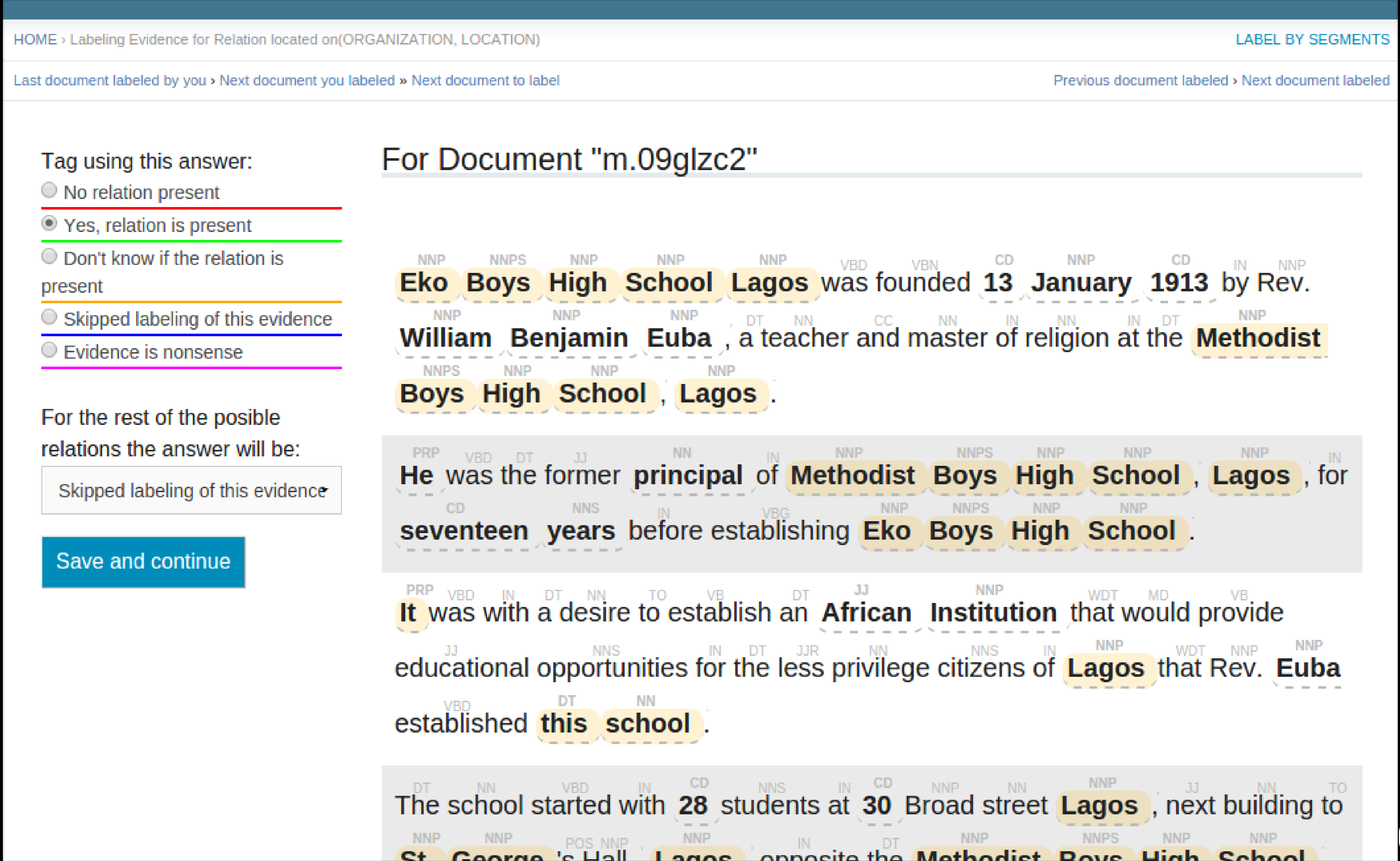
- Entity recognition
  - NER tagger (Stanford's). This one does pronoun resolution and coreferencing
  - Gazettes resolution
  - Custom rules
- Parsing (Stanford's)
- Segmentation

IEPY pipeline: Extraction

Simpler but least interesting option is **human-based**…

Allows integrating domain experts through a friendly UI

Manually entered data can feed the active learning core


HOME › Labeling Evidence for Relation located on(ORGANIZATION, LOCATION)
LABEL BY SEGMENTS
Last document labeled by you › Next document you labeled » Next document to label
Previous document labeled › Next document labeled
Save and continue
NNP NNPS NNP NNP NNP VBD VBN CD NNP CD IN NNP
NNP NNP NNP DT NN CC NN IN NN IN DT NNP
NNPS NNP NNP NNP
PRP VBD DT JJ NN IN NNP NNPS NNP NNP NNP IN
CD NNS IN VBG NNP NNPS NNP NNP
PRP VBD IN DT NN TO VB DT JJ NNP WDT MD VB
JJ NNS IN DT JJR NN NNS IN NNP WDT NNP NNP
VBD DT NN
DT NN VBD IN CD NNS IN CD NNP NN NNP JJ NN TO
NNP NNP POS NNP NNP IN DT NNP NNPS NNP NNP

Tag using this answer:

○ No relation present

● Yes, relation is present

○ Don't know if the relation is present

○ Skipped labeling of this evidence

○ Evidence is nonsense

For the rest of the posible relations the answer will be:

Skipped labeling of this evidence

For Document "m.09glzc2"

Eko Boys High School Lagos was founded 13 January 1913 by Rev. William Benjamin Euba , a teacher and master of religion at the Methodist Boys High School , Lagos .

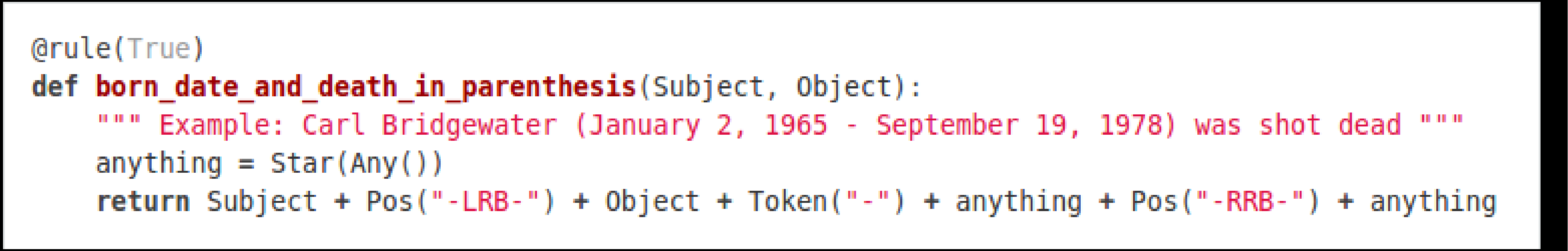
He was the former principal of Methodist Boys High School , Lagos , for seventeen years before establishing Eko Boys High School .

It was with a desire to establish an African Institution that would provide educational opportunities for the less privilege citizens of Lagos that Rev. Euba established this school .

The school started with 28 students at 30 Broad street Lagos , next building to St. George 's Hall Lagos , opposite the Methodist Boys High School

IEPY pipeline: Extraction

You can get precision (but not always high recall) with rules:



```
@rule(True)
def born_date_and_death_in_parenthesis(Subject, Object):
    """ Example: Carl Bridgewater (January 2, 1965 - September 19, 1978) was shot dead """
    anything = Star(Any())
    return Subject + Pos("-LRB-") + Object + Token("-") + anything + Pos("-RRB-") + anything
```

Rules are defined by regular expressions-like patterns on words

Some tools for debugging rules

IEPY pipeline: Extraction

You can enhance your results through active learning methods (and reduce human time required which is the end-goal)

Learns from positive and negative examples found by other methods.

Some predefined features (entity distance, bag of words, bag of bigrams, ...)

Many predefined classifiers (defaults to SVC from scikit-learn)

Tunable parameters for high precision or high recall

A real use-case: Archivo de la Memoria

Background history:

In 1976 Argentina, amidst heavy unrest and civil guerrilla/terrorism, the military forces removed the president by coup and set up a new government that lasted until 1983.

During that period, people considered “subversive” (members of the terrorist organizations, but also any political dissenter) were taken away without warrant, and many of them never seen again (killed without trial and with no record). Those are called “Desaparecidos” (disappeared)

A real use-case: Archivo de la Memoria

In many of those cases, spouses and small children of desaparecidos were also kidnapped. Also some pregnant women gave birth while imprisoned. Those children were given under clandestine adoption to families in the military or connected with them.

Today, most of the military top leaders of that age have been tried and convicted for Human Rights violations. But many people in their 30’s have been lied about their real identity. Only 119 have been found as of December of 2015. Nearly 400 haven’t been located yet.

A real use-case: Archivo de la Memoria

A collaboration work between

- The NLP group at FaMAF, UNC (a research dept)
- The Province Archives of Remembrance (http://www.apm.gov.ar/ , government agency)
    - Clearance to access classified military files
    - Recovery and conservation of documents

A real use-case: Archivo de la Memoria

Input data: Army bulletins

- Personnel movements/assignments/transfers
- Training activities (“interrogation”, “intelligence”)
- Promotions, retirements.

Scanned and OCR’d from typewritten documents. Tagged with metadata (location, date)

A real use-case: Archivo de la Memoria

Entity recognition

- Rule based: Stanford NLP is not very good in Spanish
- Rule based: To match military document conventions

A real use-case: Archivo de la Memoria

Relationship extraction

- Rule based: Taking advantage of having very regular language.
- ML based: SVM classifier to increase recall.
- Human based: some rounds of manual validation.
- ML based: retrained and relabeled after human intervention.

A real use-case: Archivo de la Memoria

Some results

- 67K people identified by rules, raised to 141K with a classifier.
- 10K transfers identified by rules, raised to 22K with a classifier

Where IEPY works well

- Single domain improves success rate
- Best on binary relations.
- Best on regular documents
- “Medium” sized datasets

Be careful about

- You are detecting occurrences, not actual entities
  - Is ambiguity or “denormalization” harmful to you?
  - For temporal events is a duplicate harmful?
- You get what the document says
  - Are your documents consistent?
  - Are your documents “truth”?

Thank you!

machinalis

Daniel F. Moisset – dmoisset@machinalis.com
tw: @dmoisset

the missing piece.