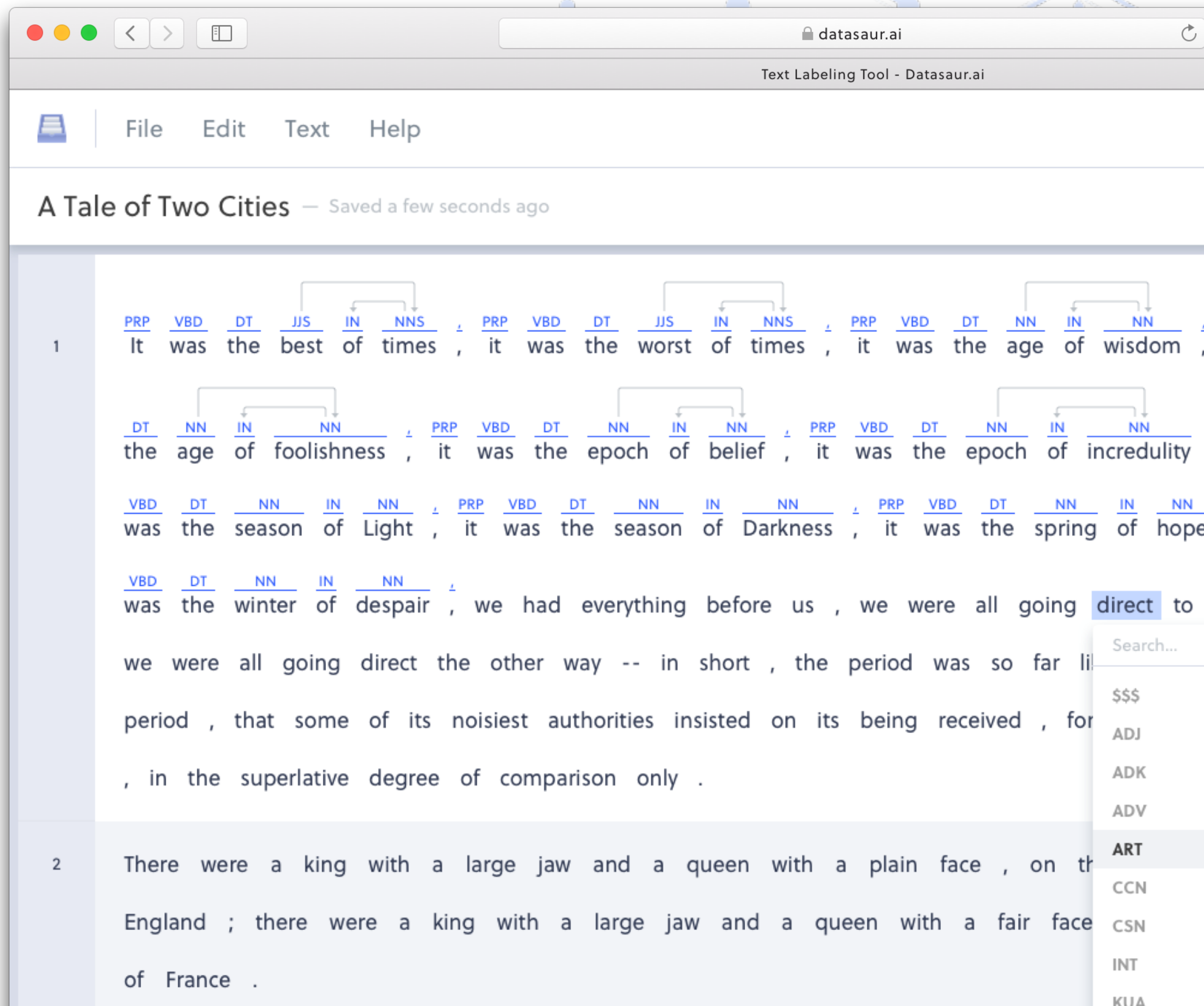
datasaur.ai

We build data labeling tools for Natural Language Processing.

February 2021

What we do

Datasaur is an **end-to-end solution for labeling data** and **using it to build and train NLP models.**


datasaur.ai
Text Labeling Tool - Datasaur.ai
File
Edit
Text
Help
PRP VBD DT JJS IN NNS , PRP VBD DT JJS IN NNS , PRP VBD DT NN IN NN
DT NN IN NN , PRP VBD DT NN IN NN , PRP VBD DT NN IN NN
VBD DT NN IN NN , PRP VBD DT NN IN NN , PRP VBD DT NN IN NN
VBD DT NN IN NN ,
Search...
$$$
ADJ
ADK
ADV
ART
CCN
CSN
INT
KUA

A Tale of Two Cities — Saved a few seconds ago

1 It was the best of times , it was the worst of times , it was the age of wisdom ,
the age of foolishness , it was the epoch of belief , it was the epoch of incredulity
was the season of Light , it was the season of Darkness , it was the spring of hope
was the winter of despair , we had everything before us , we were all going direct to
we were all going direct the other way -- in short , the period was so far li
period , that some of its noisiest authorities insisted on its being received , for
, in the superlative degree of comparison only .

2 There were a king with a large jaw and a queen with a plain face , on th
England ; there were a king with a large jaw and a queen with a fair face
of France .

About Us

- YC W20, Stanford StartX F19
- Funded by Initialized Capital, CTO of Segment, CTO of OpenAI
- 100m+ labels applied on our site annually
- Serving top companies, academic institutions, non-profits globally
- Interviewed 200+ ML teams around the world

Stanford University

StartX

Y Combinator

OpenAI

Setting the Stage 🎤

NLP use cases across industries and functions

- Product review understanding
- Social media sentiment tracking
- Forum moderation
- Customer support chatbots
- Misinformation detection in the news
- Voice assistants
- Receipt invoice and understanding
- COVID-19 research information extraction

NLP use cases across industries and functions

- **Named Entity Recognition**
  - **Classification**
- Dependency parsing
  - Part of speech
- Coreference resolution
  - **Optical character recognition**
- Aspect-based sentiment analysis

Walk… then run

Walk

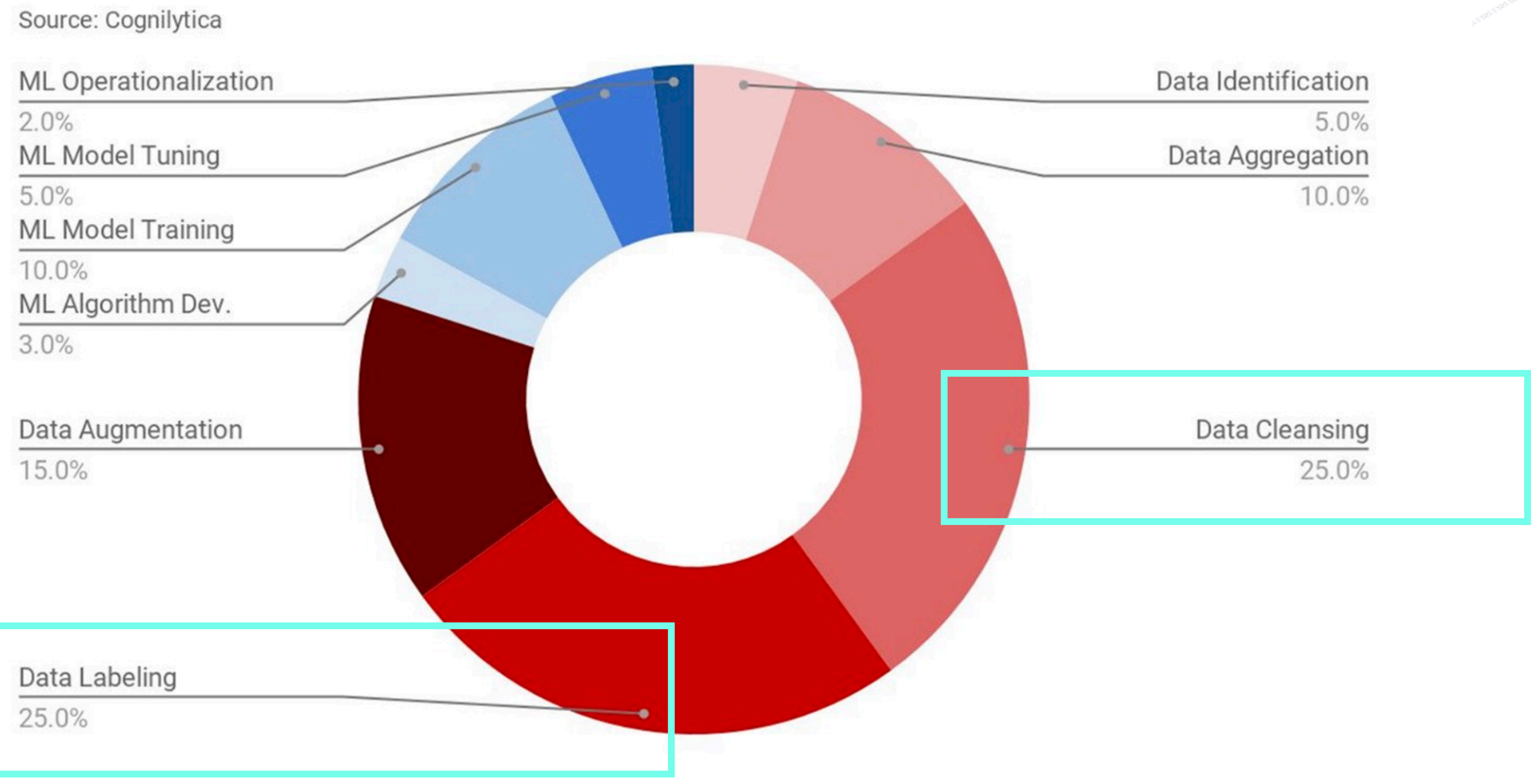
This is **Sesame Street**, a place where **people**, **birds**, and **monsters** all live in perfect harmony.	**Entity recognition**
Cookies always good. - **Positive** Mondays bad. - **Negative**	**Sentiment analysis**

Run

Bathtime can be **tiring**, but **rubber duckies** are the **best**.	**Aspect-based sentiment analysis**
Never refer to **me** as an item. **I** am a bird. -**Big Bird**	**Coreference resolution**

Data Labeling

Percentage of Time Allocated to Machine Learning Project Tasks



FACTOR DAILY

SOURCE: COGNILYTICA

Data labeling - a history

- 1963 - Cranfield Experiments
  - Experimental studies on information retrieval
- 1992 - DARPA and NIST establish Text REtrieval Conference (TREC)
  - Expand Cranfield methodology to many tracks
    - Many providers begin offering human intelligence labor for hire
- 2001 - Mechanical Turk patented
  - Named after a fake chess-playing machine
- 2004 - ESP game from CMU
  - 2 players enter labels for an image until they agree

Data labeling - a history

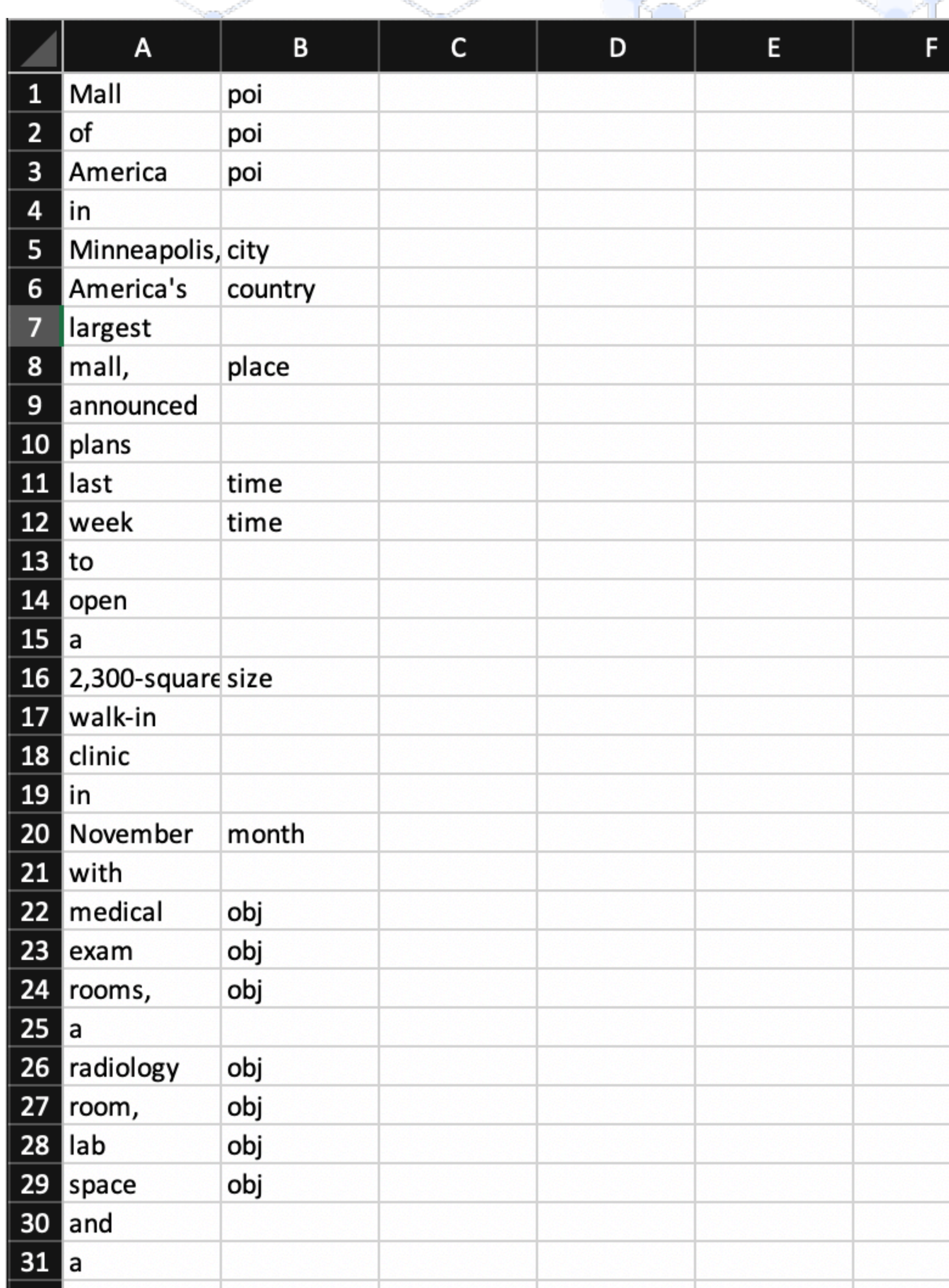
- 2005 - Amazon Mechanical Turk launch
- 2007 - CrowdFlower launch
- 2008 - Samasource
- 2010 - Human Computation Conference launch
- 2012 - iMerit
- 2016 - Scale AI
- 2019 - **Datasaur**

The Business💰

Universal problem for ML Engineers

A lack of standardization and experience in the rapidly growing AI space has led to:

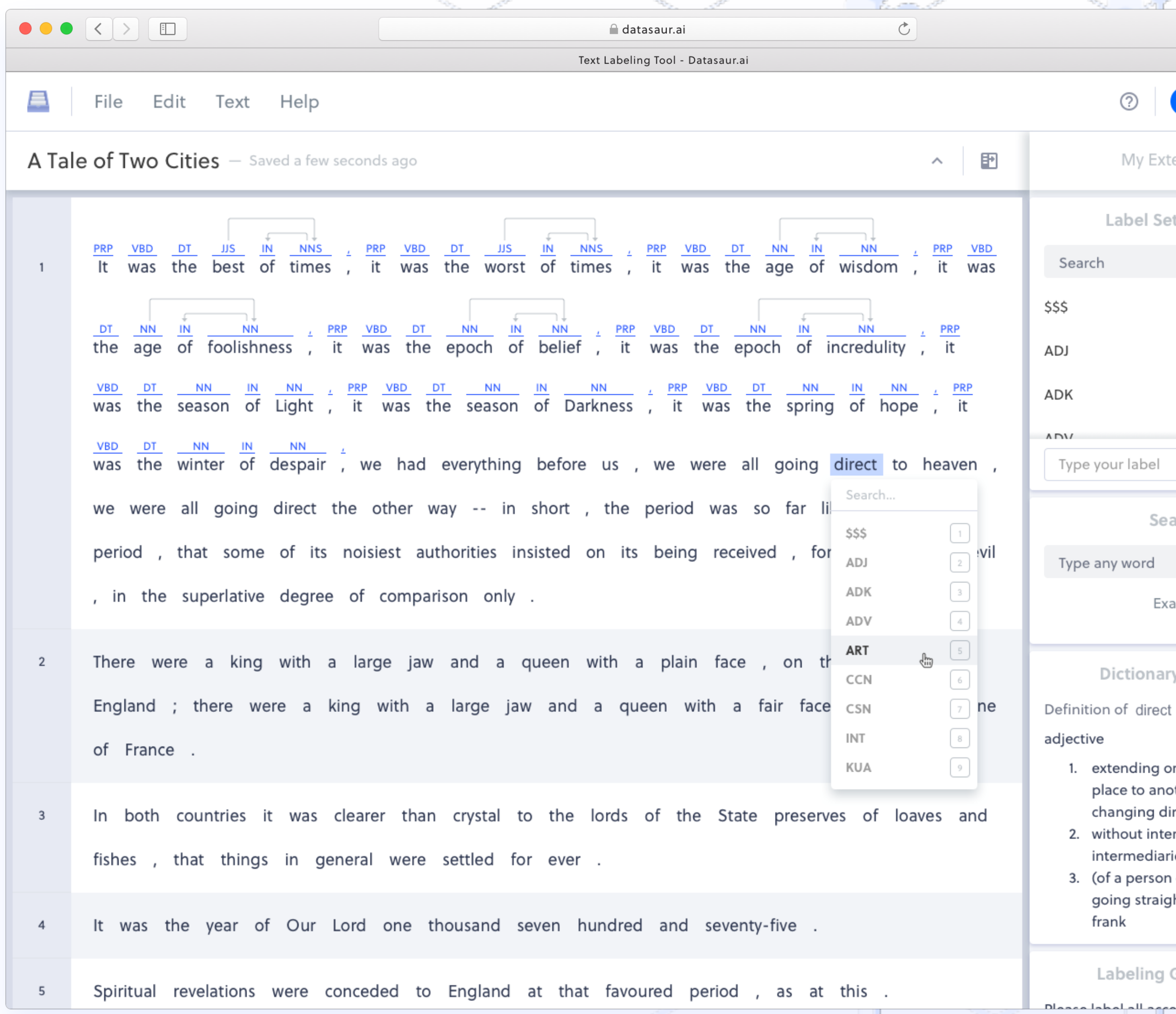
- Painstakingly manual labeling
- Error-prone workflows
- Improper handling of sensitive data
- Inadequate security
- Limited file format support
- Lack of visibility into labeler performance
- Team scalability issues
- Adoption of multiple, segregated tools



	A	B	C	D	E	F
1	Mall	poi				
2	of	poi				
3	America	poi				
4	in					
5	Minneapolis,	city				
6	America's	country				
7	largest					
8	mall,	place				
9	announced					
10	plans					
11	last	time				
12	week	time				
13	to					
14	open					
15	a					
16	2,300-square	size				
17	walk-in					
18	clinic					
19	in					
20	November	month				
21	with					
22	medical	obj				
23	exam	obj				
24	rooms,	obj				
25	a					
26	radiology	obj				
27	room,	obj				
28	lab	obj				
29	space	obj				
30	and					
31	a					

Datasaur.ai

- **Over 100 million** labels applied annually
- Community of top NLP experts globally
- **ML-assisted**, semi-automated labeling
- Built from the ground up for power users
- Choice of **on-cloud and on-premise** deployment
- Comprehensive API for end-to-end integration
- Military-grade security
- Comprehensive all-in-one tool for labeling and workforce management


datasaur.ai
Text Labeling Tool - Datasaur.ai
File
Edit
Text
Help
My Exte
Label Se
Search
$$$
ADJ
ADK
ADV
Type your label
Sea
Type any word
Exa
Dictionar
Search...
$$$ 1
ADJ 2
ADK 3
ADV 4
ART 5
CCN 6
CSN 7
INT 8
KUA 9
Labeling G
PRP VBD DT JJS IN NNS , PRP VBD DT JJS IN NNS , PRP VBD DT NN IN NN , PRP VBD
DT NN IN NN , PRP VBD DT NN IN NN , PRP VBD DT NN IN NN , PRP
VBD DT NN IN NN , PRP VBD DT NN IN NN , PRP VBD DT NN IN NN , PRP
VBD DT NN IN NN ,

A Tale of Two Cities — Saved a few seconds ago

1 It was the best of times , it was the worst of times , it was the age of wisdom , it was the age of foolishness , it was the epoch of belief , it was the epoch of incredulity , it was the season of Light , it was the season of Darkness , it was the spring of hope , it was the winter of despair , we had everything before us , we were all going direct to heaven , we were all going direct the other way -- in short , the period was so far li period , that some of its noisiest authorities insisted on its being received , for vil , in the superlative degree of comparison only .

2 There were a king with a large jaw and a queen with a plain face , on th England ; there were a king with a large jaw and a queen with a fair face ne of France .

3 In both countries it was clearer than crystal to the lords of the State preserves of loaves and fishes , that things in general were settled for ever .

4 It was the year of Our Lord one thousand seven hundred and seventy-five .

5 Spiritual revelations were conceded to England at that favoured period , as at this .

Definition of direct

adjective

1. extending or place to ano changing dir
2. without inter intermediari
3. (of a person going straigl frank

Considering a labeling solution?

Don’t reinvent the wheel

- Many companies underestimate the complexities of building a labeling interface. NLP projects often require multiple forms of labeling projects (e.g. named entity recognition, document classification, sentiment analysis).

Existing solutions lack key functionality

- In addition to providing a comprehensive interface, your team will also require features such as label review and spot-checking to ensure the quality of the labeling output.

Your team’s needs will scale as it grows

- As your labeling efforts mature, you will need the ability to delegate, set priorities and deadlines, and measure the performance of individual labelers.

Deployment flexibility

Sensitive information?

- Many projects involve private or proprietary data. For compliance reasons or competitive reasons, it is not an acceptable option to upload the data to cloud solutions or outsource to third-party vendors. Datasaur provides on-premise / hybrid options.

Working with subject-matter experts?

- Some teams pay experts upwards of $500 per hour to label their data. Each minute is costly for such projects. Datasaur's solutions have been proven to increase efficiency by more than 100%. Semi-automated labeling and dynamic project delegation can help keep costs low.

Datasaur Business Value

Save on engineering resources

- Deploy engineering efforts to projects that truly matter for your company. Save time re-building and updating an existing solution.

Improve productivity and quality

- Equip your team with the best labeling solution in the market. Improve productivity by over 100%. Utilize industry best-practices to ensure data quality.

Minimize risk: security and compliance

- Datasaur's military-grade security ensures your data is in the right hands at all times.

Datasaur for teams of all sizes

Growth

- Teams of up to 10
- 100k labels monthly
- Unlimited storage
- Workforce management
- Data validation and review
- Full data privacy guaranteed

Enterprise

- Everything in Growth and…
- 1m labels monthly
- On-prem / hybrid deployment options
- Data import/export via API
- Dedicated support with 24-hour SLA
- $500 installation fee waived if paying for 12 months upfront

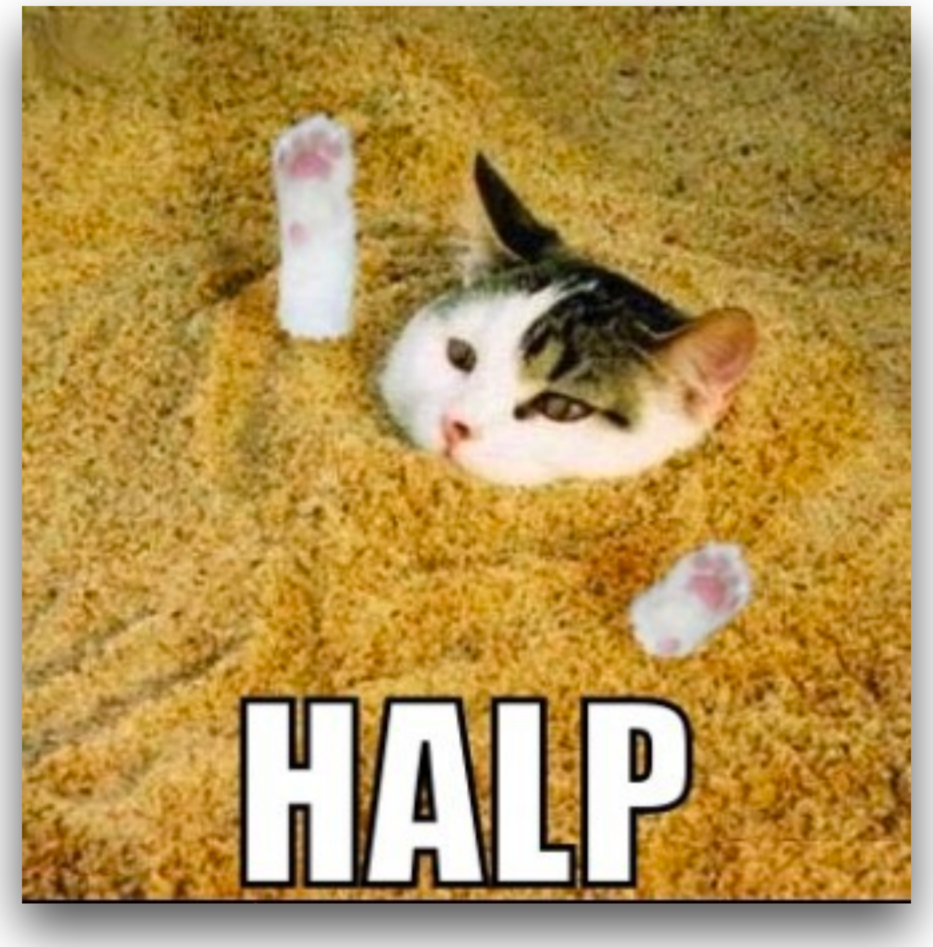
For all tiers

- Optimized Labeling interface
- Built-in Intelligence
- Regular expression extension
- Access to Datasaur Platform Extensions

Practical labeling

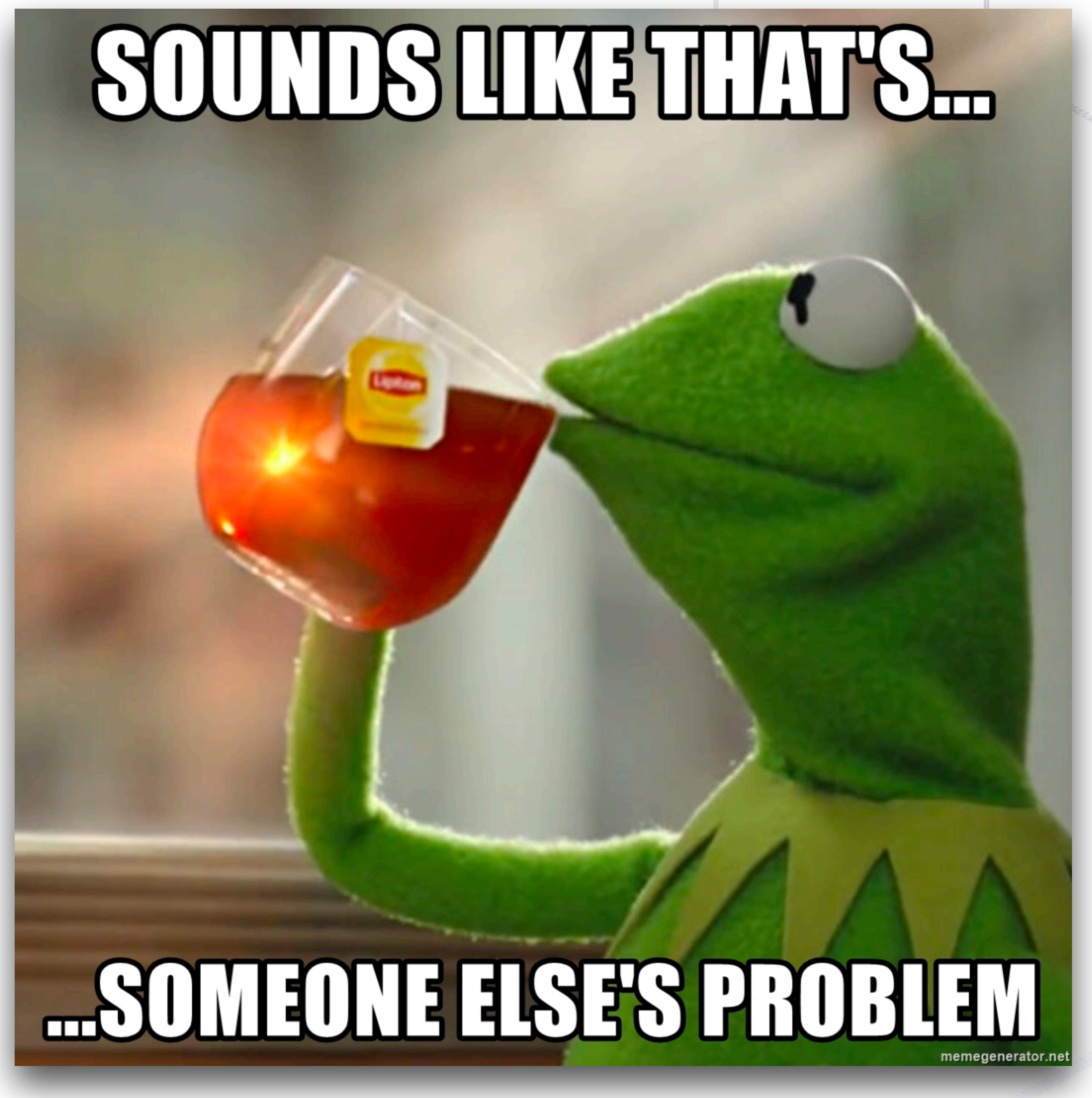
The "Problem"

- The data needs labels
- LOTS of labels
- Millions of labels by the end of next week
- There's only 3 of us



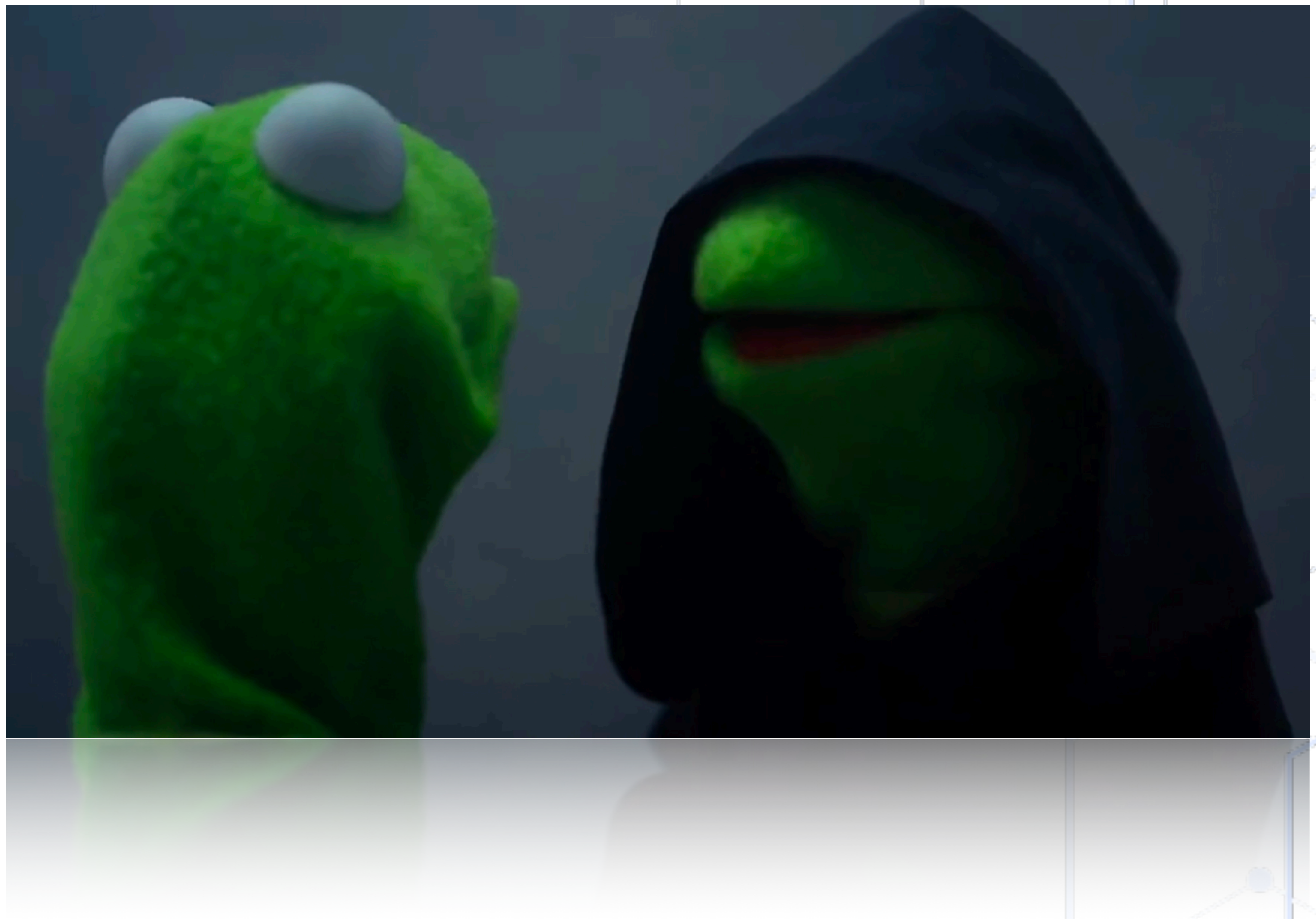
The "Solution"

- Labeling experts
- Lower cost labor
- Easily scalable



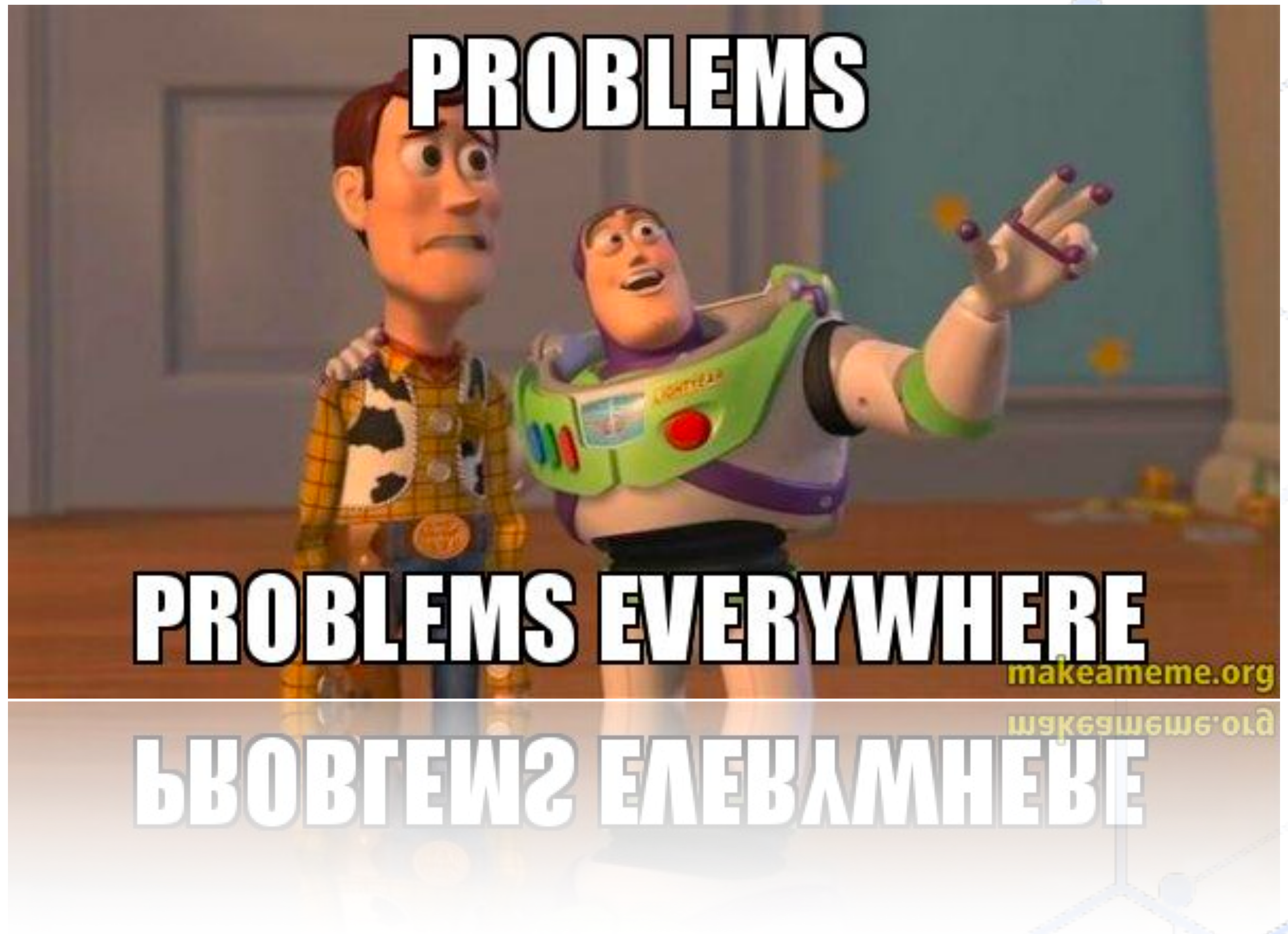
Reality

- Training takes weeks
- Guidelines are 20+ pages
- High turnover rate



The Problems

1. Crowd-sourced vs. managed services vs. internal
2. Quality vs. timeline vs. budget
3. Domain expertise
4. Ethical sourcing
5. Data privacy and compliance
6. Choosing the right tools
7. Designing the right job



1. Crowdsourced vs. managed service vs. internal

2. Quality vs. timeline vs. budget

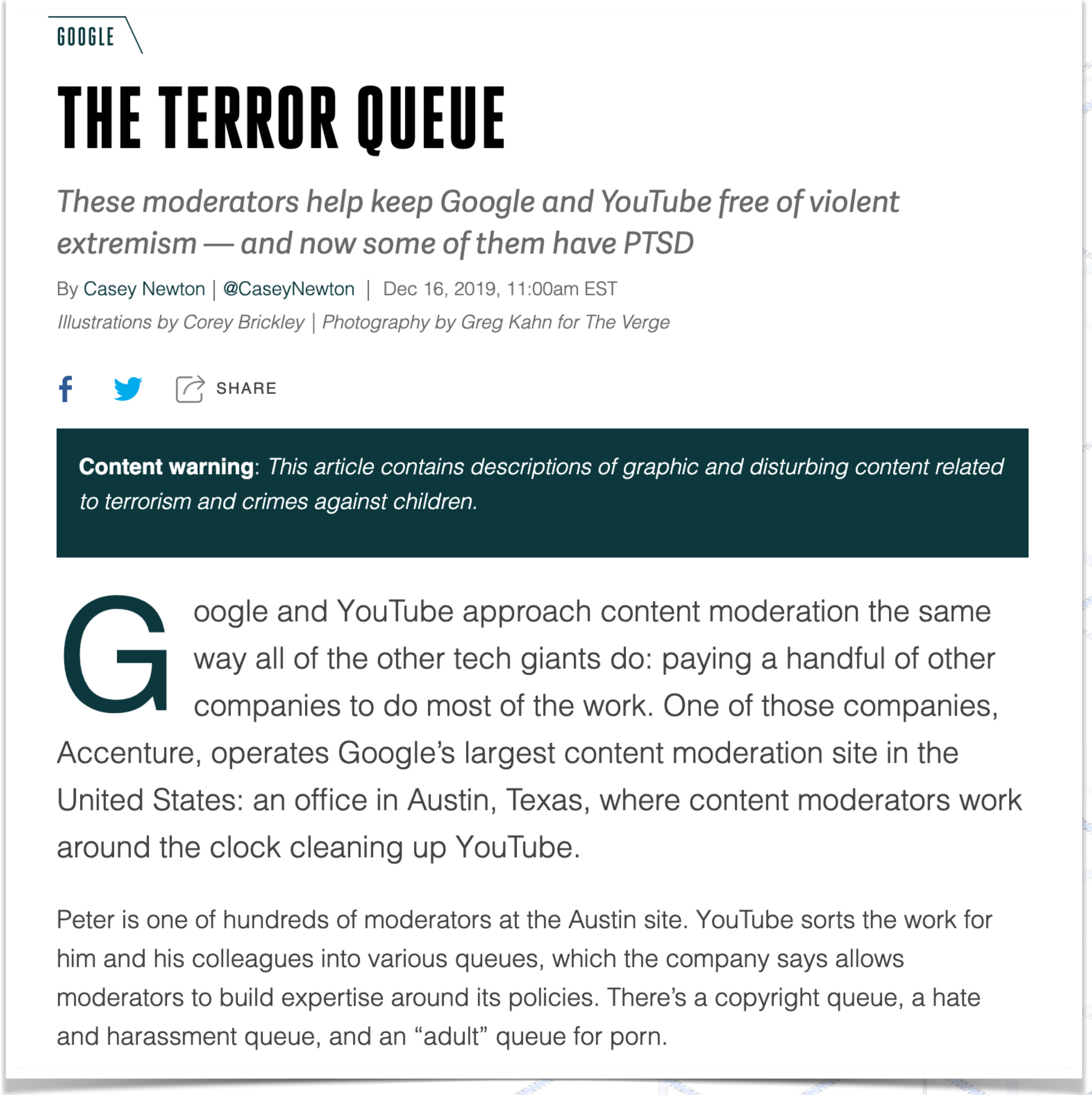
3. Domain expertise

- What level of quality is required?
- How complex is the job?
- How much training is required?
- How much is your company willing to spend?
- How much labeling is required?
    - By what deadline?
    - At what frequency?
    - With what SLA?

4. Ethical Sourcing

(This is a business problem, in addition to an ethical one)

- Who is doing your labeling?
- Under what conditions?
- For how many hours?
- Do they get paid benefits?
- Who provides the equipment?
- Is training provided?
- Are they full-time or contractors?



GOOGLE

THE TERROR QUEUE

These moderators help keep Google and YouTube free of violent extremism — and now some of them have PTSD

By Casey Newton | @CaseyNewton | Dec 16, 2019, 11:00am EST

Illustrations by Corey Brickley | Photography by Greg Kahn for The Verge

SHARE

Content warning: *This article contains descriptions of graphic and disturbing content related to terrorism and crimes against children.*

Google and YouTube approach content moderation the same way all of the other tech giants do: paying a handful of other companies to do most of the work. One of those companies, Accenture, operates Google's largest content moderation site in the United States: an office in Austin, Texas, where content moderators work around the clock cleaning up YouTube.

Peter is one of hundreds of moderators at the Austin site. YouTube sorts the work for him and his colleagues into various queues, which the company says allows moderators to build expertise around its policies. There's a copyright queue, a hate and harassment queue, and an "adult" queue for porn.

4. Ethical Sourcing

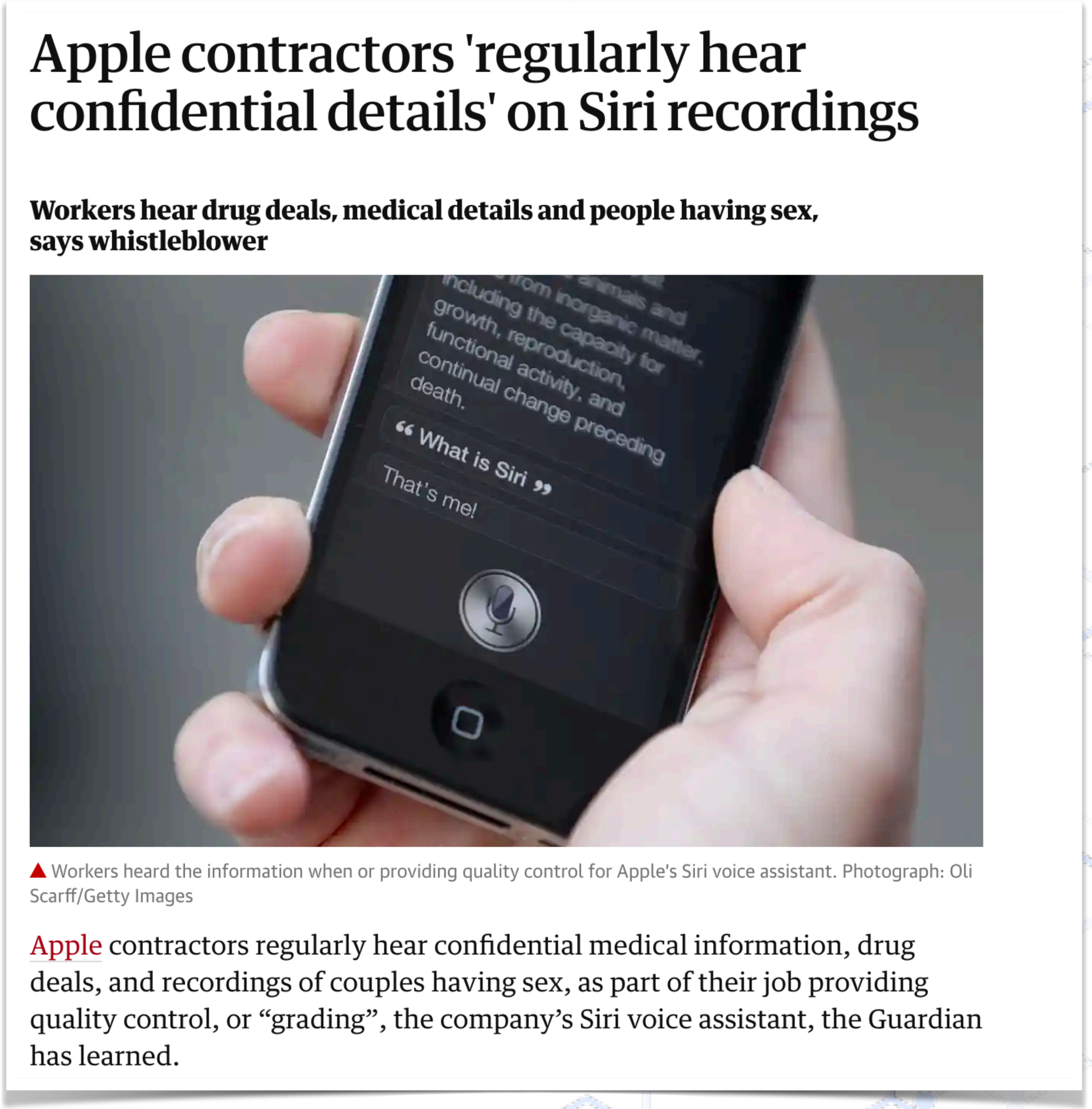
A Labeler's POV

- Tedious, boring, repetitive work
- Constantly being evaluated / fearful for job security
- Rules of the game constantly changing
- Building the system that will automate their job away
- Slow/poor escalation system
- Faceless Corporate Boss
- Need to learn to "game" the system

Corporate needs you to find the difference between this picture and this picture
They're the same picture

5. Data Privacy and Compliance

- Who can access your data?
- What are the regulatory requirements?
- How would you handle a data leak?



Apple contractors 'regularly hear confidential details' on Siri recordings

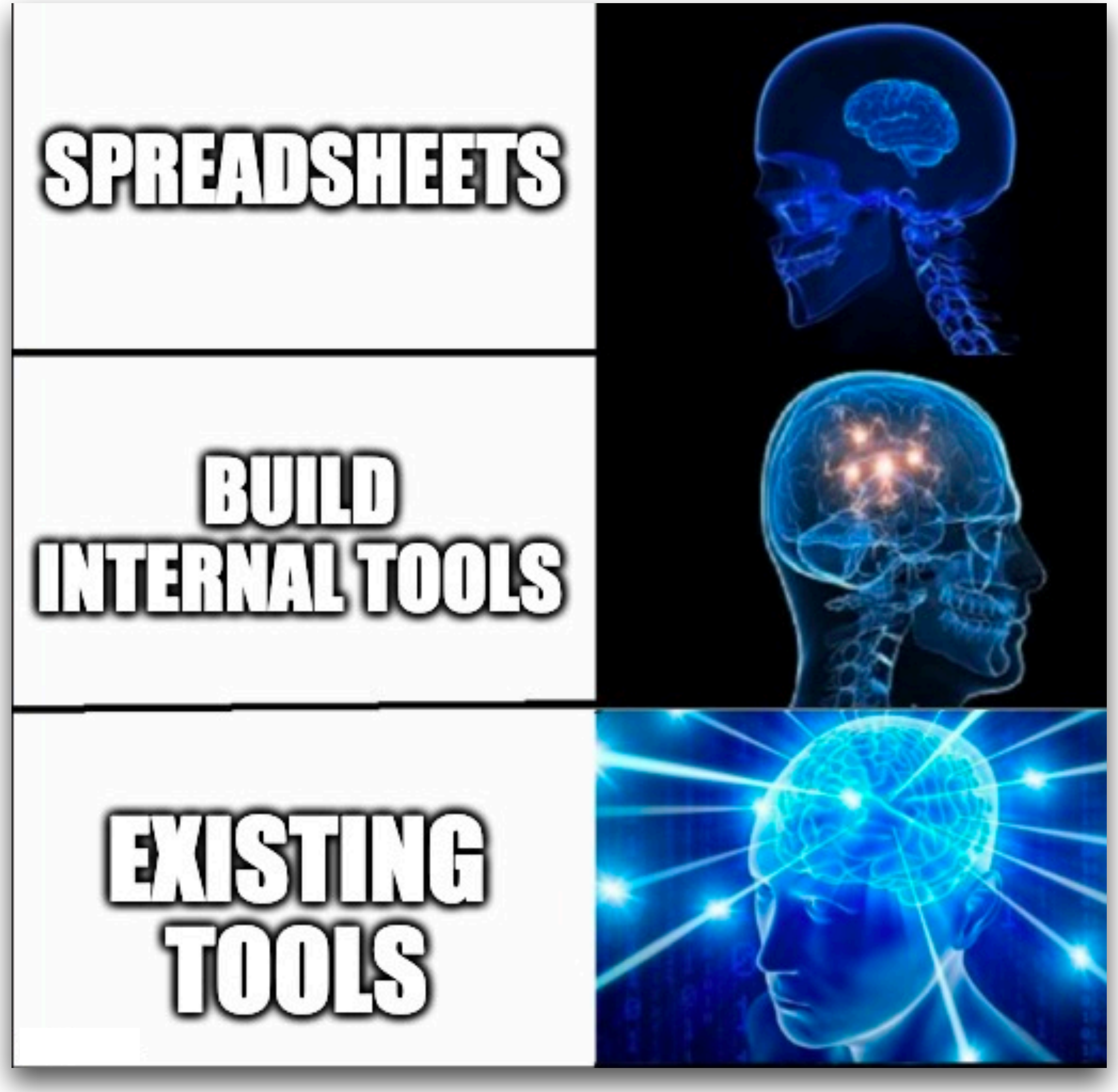
Workers hear drug deals, medical details and people having sex, says whistleblower

including the capacity for growth, reproduction, functional activity, and continual change preceding death.
“What is Siri”
That's me!

▲ Workers heard the information when or providing quality control for Apple's Siri voice assistant. Photograph: Oli Scarff/Getty Images

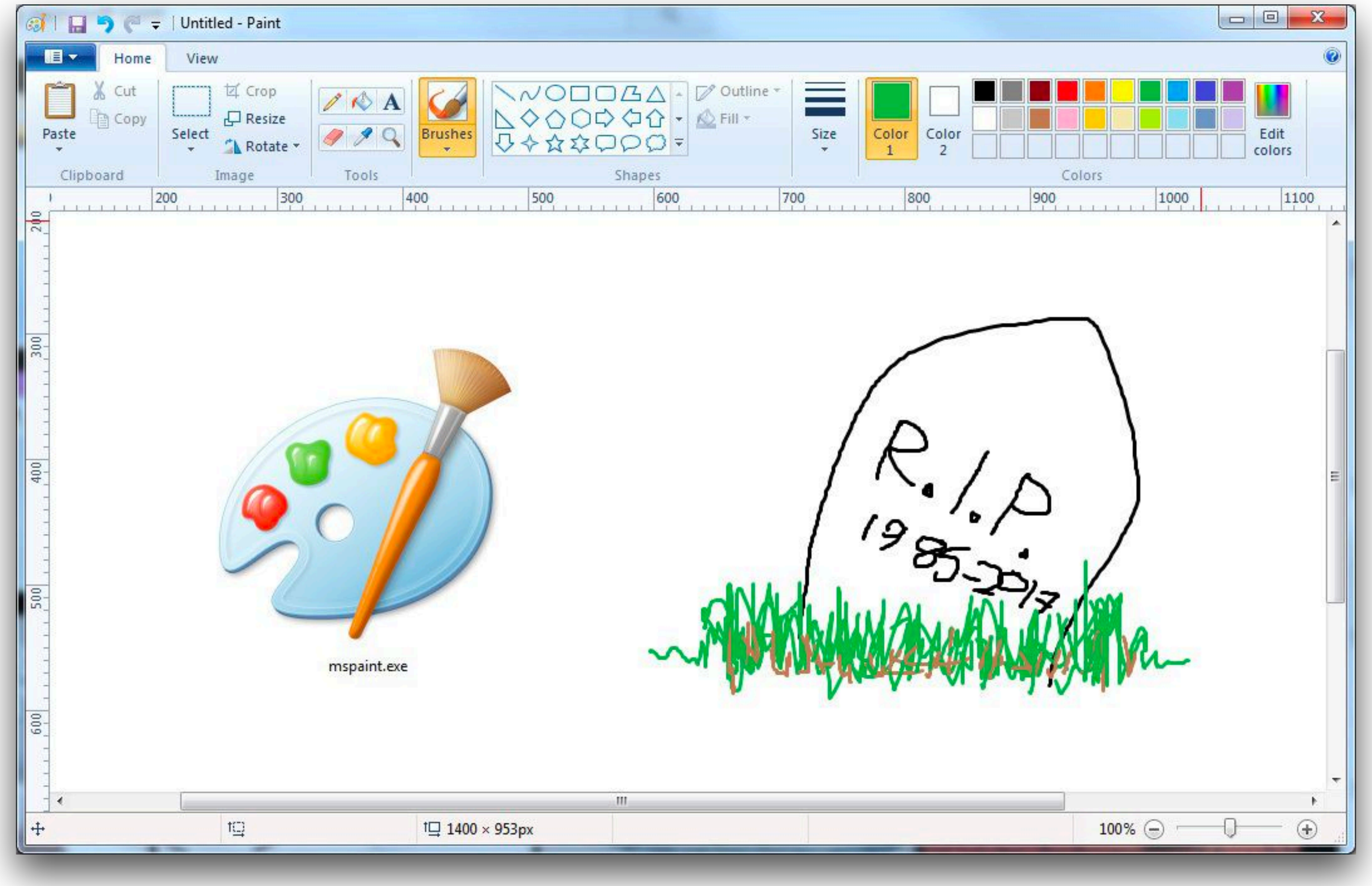
Apple contractors regularly hear confidential medical information, drug deals, and recordings of couples having sex, as part of their job providing quality control, or “grading”, the company’s Siri voice assistant, the Guardian has learned.

6. Choosing the right tools

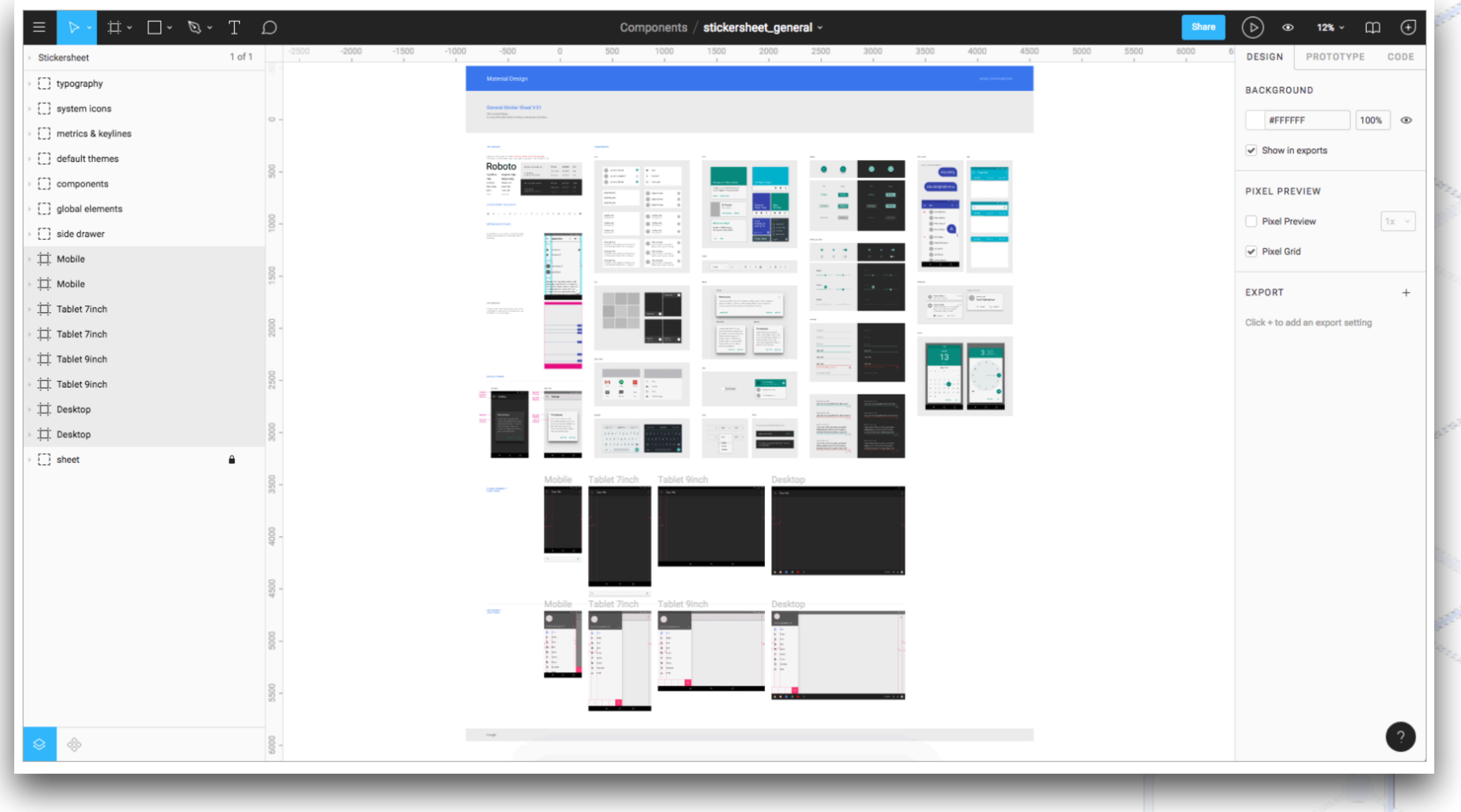


- Garbage in, garbage out
- Budget
- Amount of data
- Number of labelers
- Complexity of task
- Data privacy/compliance

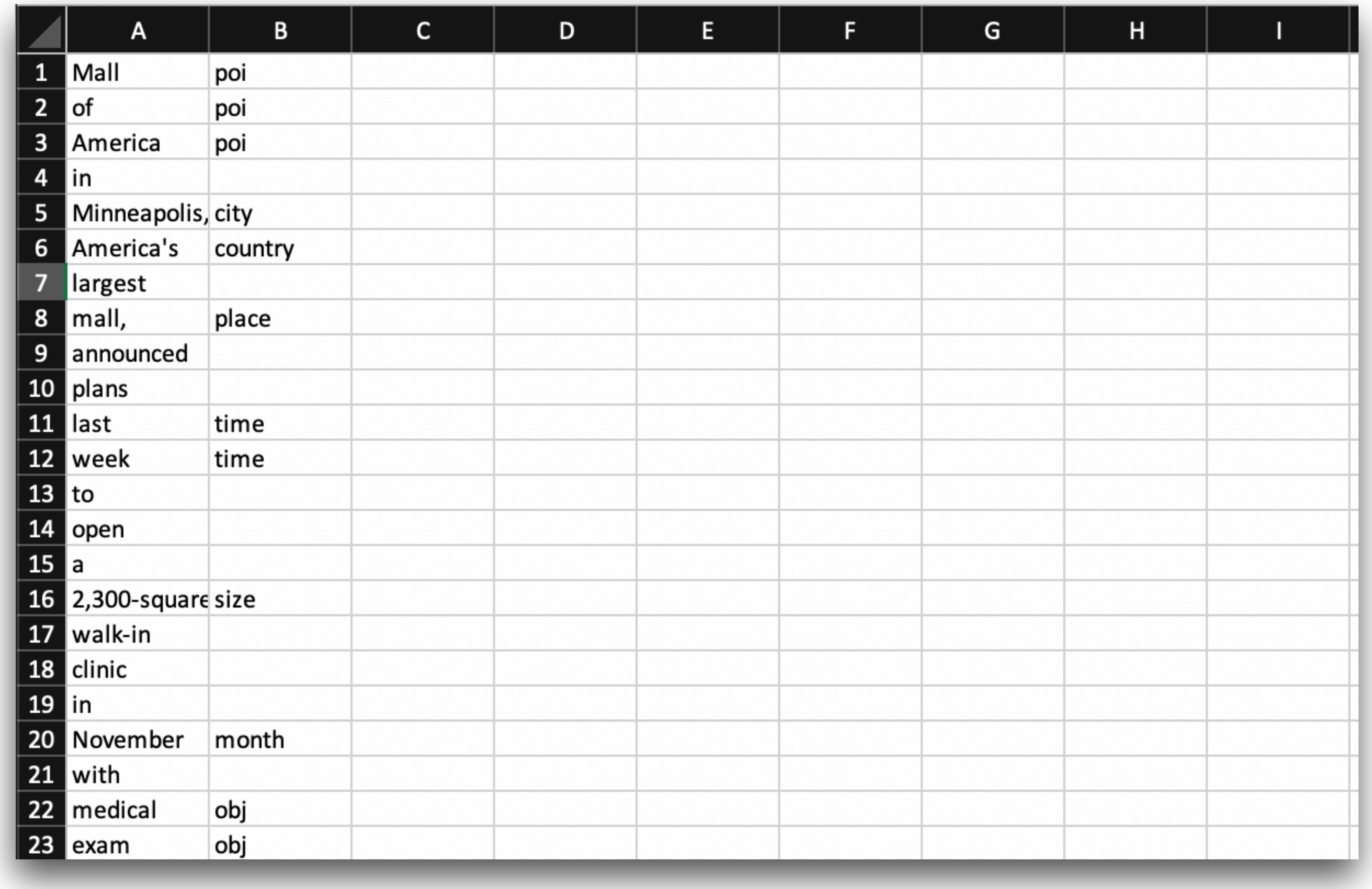
6. Choosing the right tools



is to

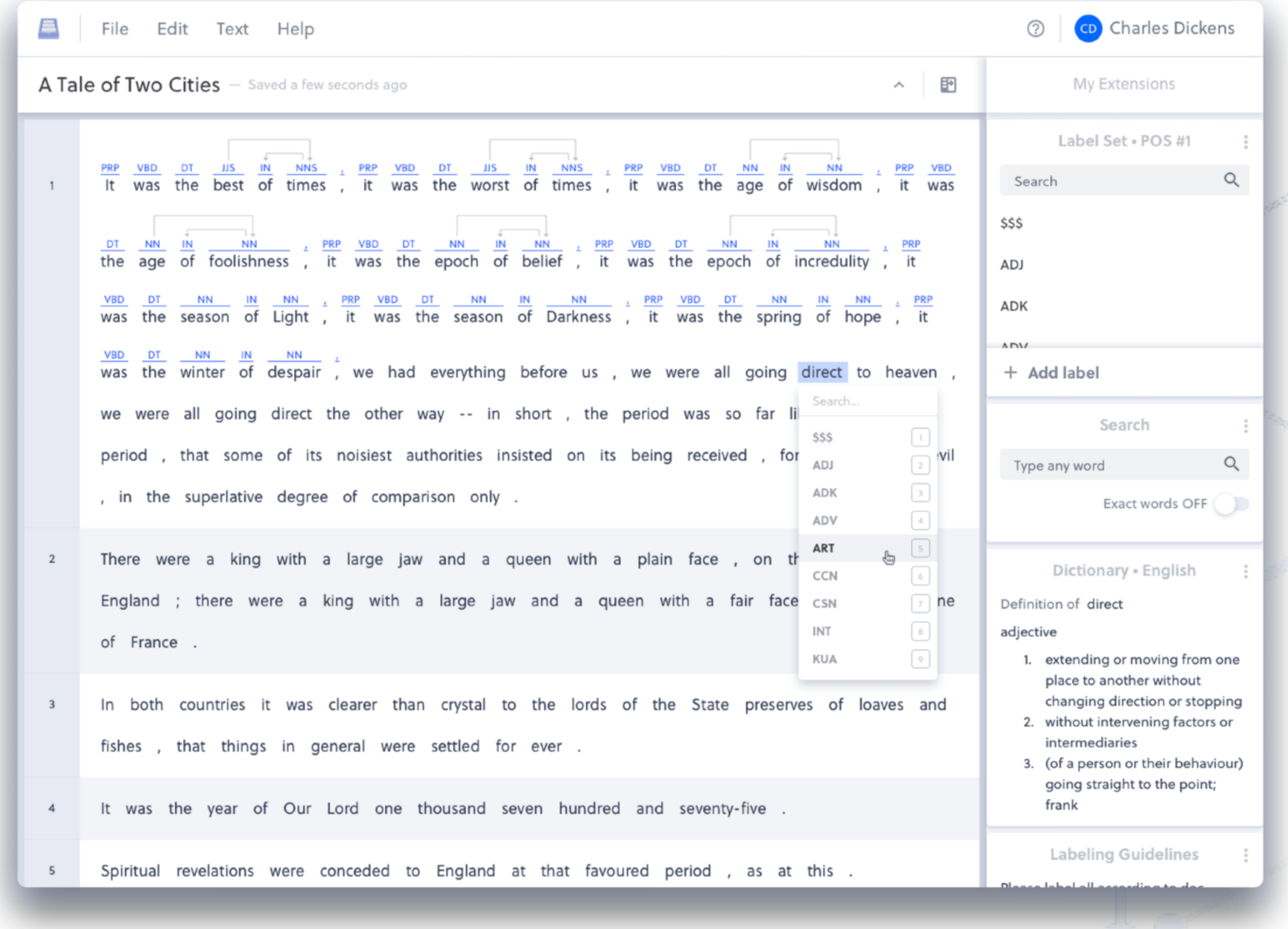


6. Choosing the right tools



	A	B	C	D	E	F	G	H	I
1	Mall	poi							
2	of	poi							
3	America	poi							
4	in								
5	Minneapolis,	city							
6	America's	country							
7	largest								
8	mall,	place							
9	announced								
10	plans								
11	last	time							
12	week	time							
13	to								
14	open								
15	a								
16	2,300-square	size							
17	walk-in								
18	clinic								
19	in								
20	November	month							
21	with								
22	medical	obj							
23	exam	obj							

is to



File
Edit
Text
Help
Charles Dickens
My Extensions
Label Set • POS #1
Search
$$$
ADJ
ADK
+ Add label
Search
Type any word
Exact words OFF
Search...
$$$ 1
ADJ 2
ADK 3
ADV 4
ART 5
CCN 6
CSN 7
INT 8
KUA 9
PRP VBD DT JJS IN NNS , PRP VBD DT JJS IN NNS , PRP VBD DT NN IN NN , PRP VBD
DT NN IN NN , PRP VBD DT NN IN NN , PRP VBD DT NN IN NN , PRP
VBD DT NN IN NN , PRP VBD DT NN IN NN , PRP VBD DT NN IN NN , PRP
VBD DT NN IN NN ,

A Tale of Two Cities — Saved a few seconds ago

1 It was the best of times , it was the worst of times , it was the age of wisdom , it was the age of foolishness , it was the epoch of belief , it was the epoch of incredulity , it was the season of Light , it was the season of Darkness , it was the spring of hope , it was the winter of despair , we had everything before us , we were all going direct to heaven , we were all going direct the other way -- in short , the period was so far li vil , in the superlative degree of comparison only .

2 There were a king with a large jaw and a queen with a plain face , on th England ; there were a king with a large jaw and a queen with a fair face ne of France .

3 In both countries it was clearer than crystal to the lords of the State preserves of loaves and fishes , that things in general were settled for ever .

4 It was the year of Our Lord one thousand seven hundred and seventy-five .

5 Spiritual revelations were conceded to England at that favoured period , as at this .

Dictionary • English

Definition of direct

adjective

1. extending or moving from one place to another without changing direction or stopping
2. without intervening factors or intermediaries
3. (of a person or their behaviour) going straight to the point; frank

Labeling Guidelines

7. Designing the Right Job

- If the job could be done through algorithms or rules, why do you need a human?
- How do you account for human [error, bias, fatigue]?
- Are your labelers messing up, or is the task ill-defined?
- Should they channel the end user or channel your product vision?
- How much time should be spent per task?

#🐭🐰

7. Designing the Right Job

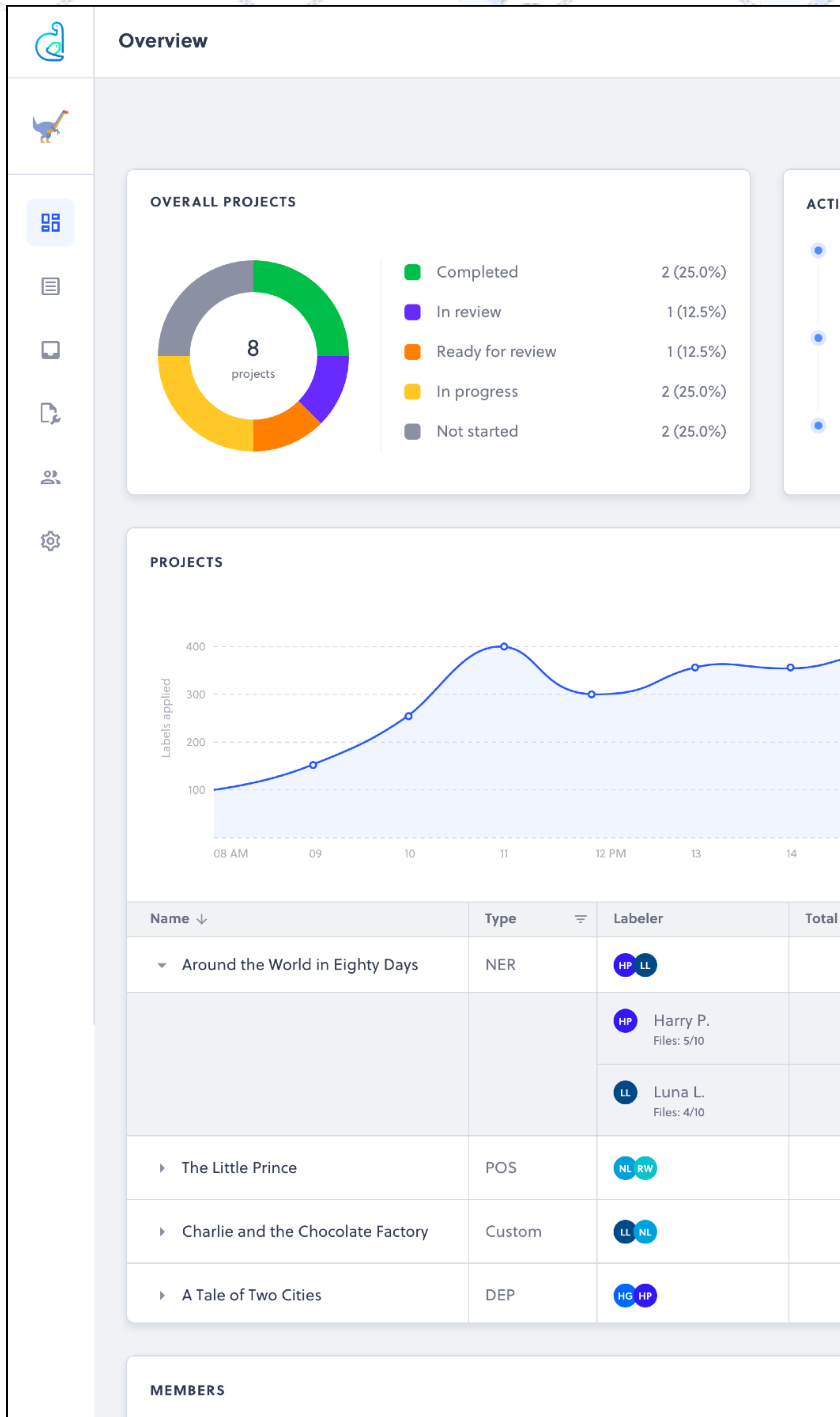
"It's not possible for an individual employee or group of employees to manipulate our search results," *Pichai said.* ***"I lead this company without political bias and work to ensure that our products continue to operate that way."*** *He reiterated that Google's search results can't be manipulated by one person and that it's designed to be robust and responsive to user feedback.*

Google Testimony before Congress

Datasaur 🦕

What we do

- Industry's best **interface** for human-in-the-loop text labeling
- Under-the-hood **intelligence** for semi-automation
- **Workforce management** platform
- **API-friendly** for direct integration
- Secure **end-to-end encryption** with data privacy at the forefront



Overview

OVERALL PROJECTS

8 projects

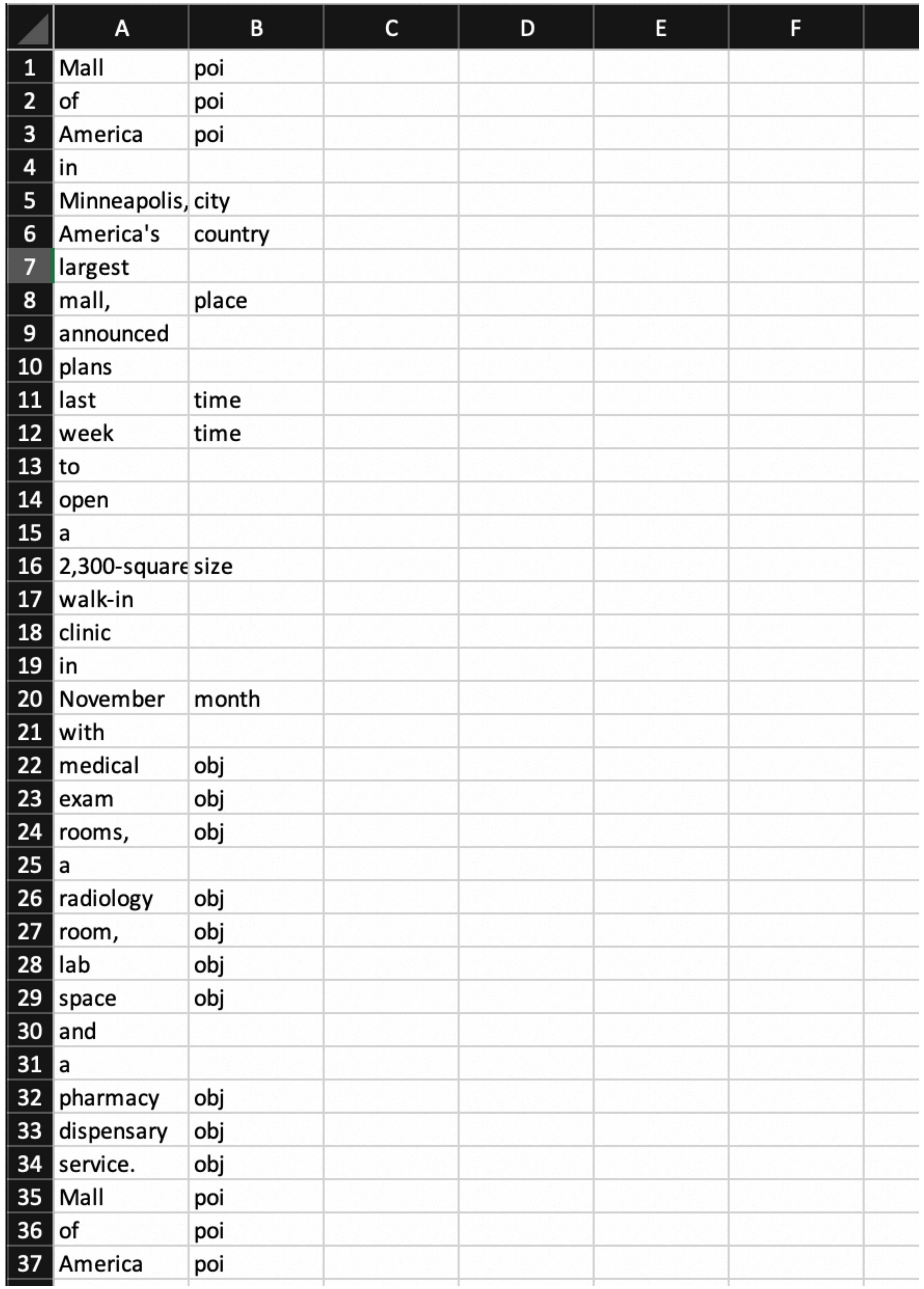
🟩	Completed	2 (25.0%)
🟪	In review	1 (12.5%)
🟧	Ready for review	1 (12.5%)
🟨	In progress	2 (25.0%)
⬜	Not started	2 (25.0%)

ACTI

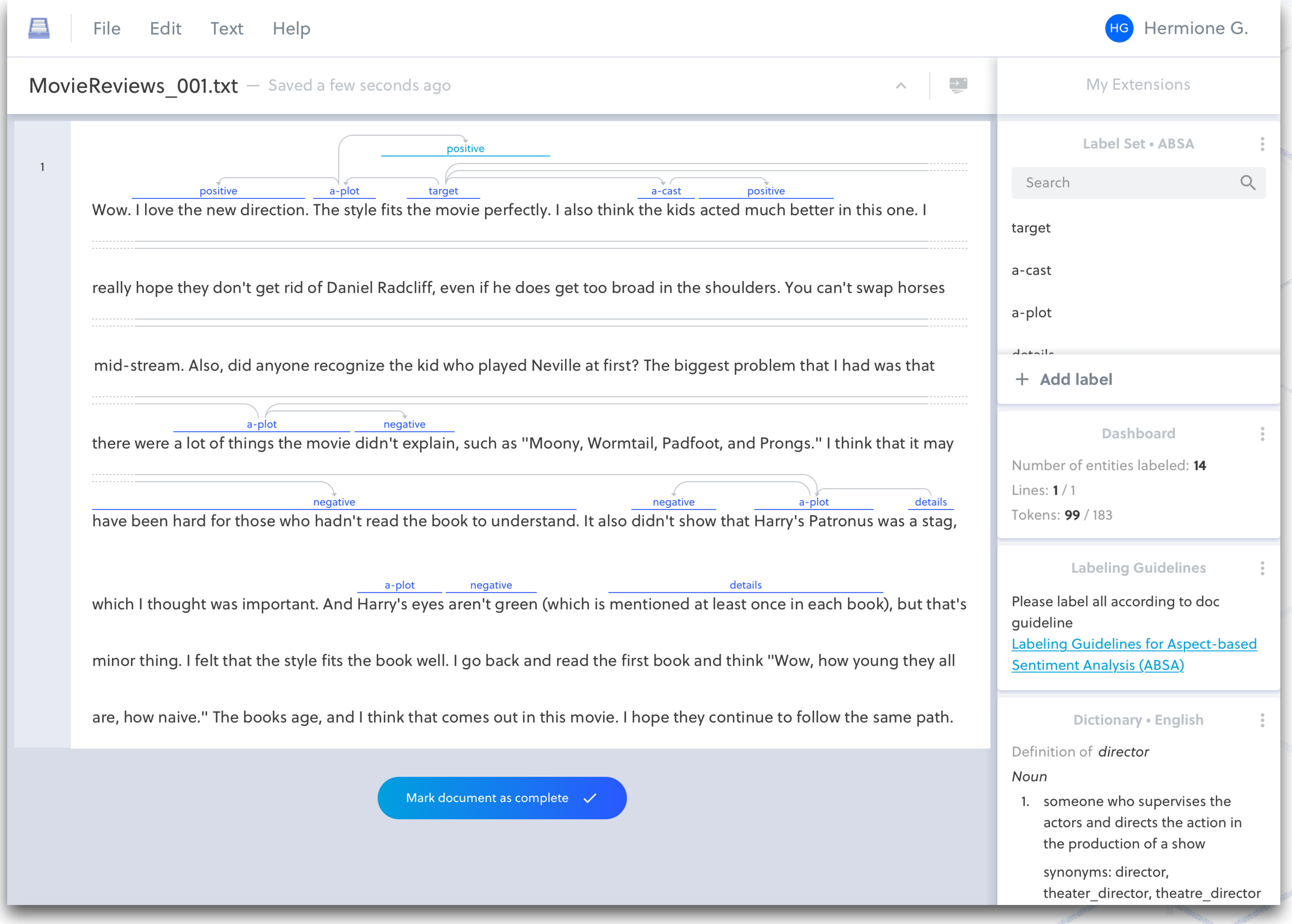
PROJECTS

Name ↓	Type	Labeler	Total
Around the World in Eighty Days	NER	HP LL	
		HP Harry P. Files: 5/10	
		LL Luna L. Files: 4/10	
The Little Prince	POS	NL RW	
Charlie and the Chocolate Factory	Custom	LL NL	
A Tale of Two Cities	DEP	HG HP	

MEMBERS



	A	B	C	D	E	F
1	Mall	poi				
2	of	poi				
3	America	poi				
4	in					
5	Minneapolis,	city				
6	America's	country				
7	largest					
8	mall,	place				
9	announced					
10	plans					
11	last	time				
12	week	time				
13	to					
14	open					
15	a					
16	2,300-square	size				
17	walk-in					
18	clinic					
19	in					
20	November	month				
21	with					
22	medical	obj				
23	exam	obj				
24	rooms,	obj				
25	a					
26	radiology	obj				
27	room,	obj				
28	lab	obj				
29	space	obj				
30	and					
31	a					
32	pharmacy	obj				
33	dispensary	obj				
34	service.	obj				
35	Mall	poi				
36	of	poi				
37	America	poi				



MovieReviews_001.txt — Saved a few seconds ago

1 Wow. I love the new direction. The style fits the movie perfectly. I also think the kids acted much better in this one. I really hope they don't get rid of Daniel Radcliff, even if he does get too broad in the shoulders. You can't swap horses mid-stream. Also, did anyone recognize the kid who played Neville at first? The biggest problem that I had was that there were a lot of things the movie didn't explain, such as "Moony, Wormtail, Padfoot, and Prongs." I think that it may have been hard for those who hadn't read the book to understand. It also didn't show that Harry's Patronus was a stag, which I thought was important. And Harry's eyes aren't green (which is mentioned at least once in each book), but that's minor thing. I felt that the style fits the book well. I go back and read the first book and think "Wow, how young they all are, how naive." The books age, and I think that comes out in this movie. I hope they continue to follow the same path.

Number of entities labeled: **14**

Lines: **1** / 1

Tokens: **99** / 183

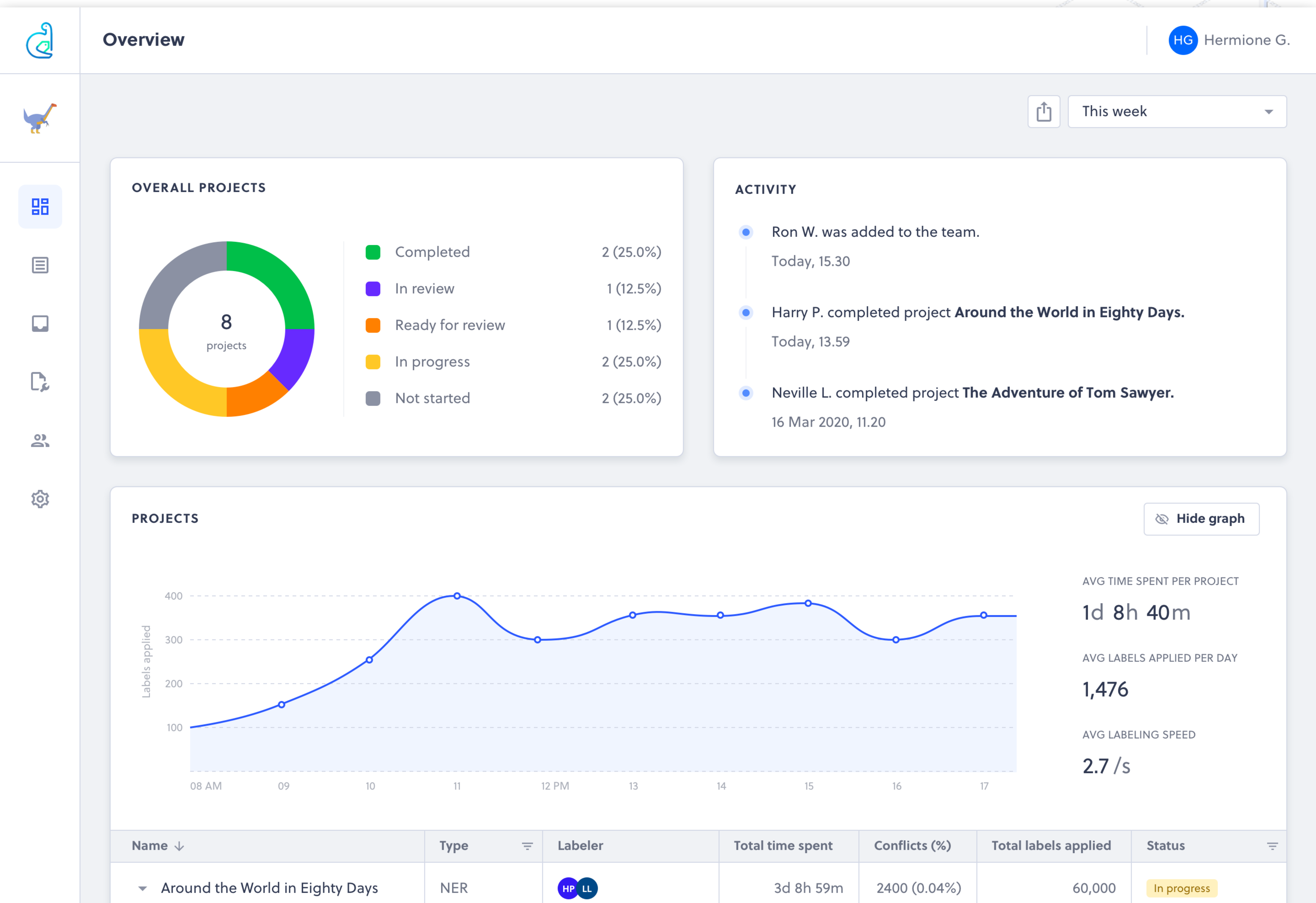
Please label all according to doc guideline
Labeling Guidelines for Aspect-based Sentiment Analysis (ABSA)

Definition of *director*

Noun

1. someone who supervises the actors and directs the action in the production of a show

   synonyms: director, theater_director, theatre_director



OVERALL PROJECTS

8 projects

🟩	Completed	2 (25.0%)
🟪	In review	1 (12.5%)
🟧	Ready for review	1 (12.5%)
🟨	In progress	2 (25.0%)
▨	Not started	2 (25.0%)

ACTIVITY

- Ron W. was added to the team.
  Today, 15.30
- Harry P. completed project **Around the World in Eighty Days.**
  Today, 13.59
- Neville L. completed project **The Adventure of Tom Sawyer.**
  16 Mar 2020, 11.20

PROJECTS

AVG TIME SPENT PER PROJECT
1d 8h 40m

AVG LABELS APPLIED PER DAY
1,476

AVG LABELING SPEED
2.7 /s

Name ↓	Type	Labeler	Total time spent	Conflicts (%)	Total labels applied	Status
Around the World in Eighty Days	NER	HP LL	3d 8h 59m	2400 (0.04%)	60,000	In progress

datasaur.ai/sign-up

ivan@datasaur.ai

https://www.linkedin.com/in/iylee/

Thank you

ivan@datasaur.ai

https://www.linkedin.com/in/iylee/