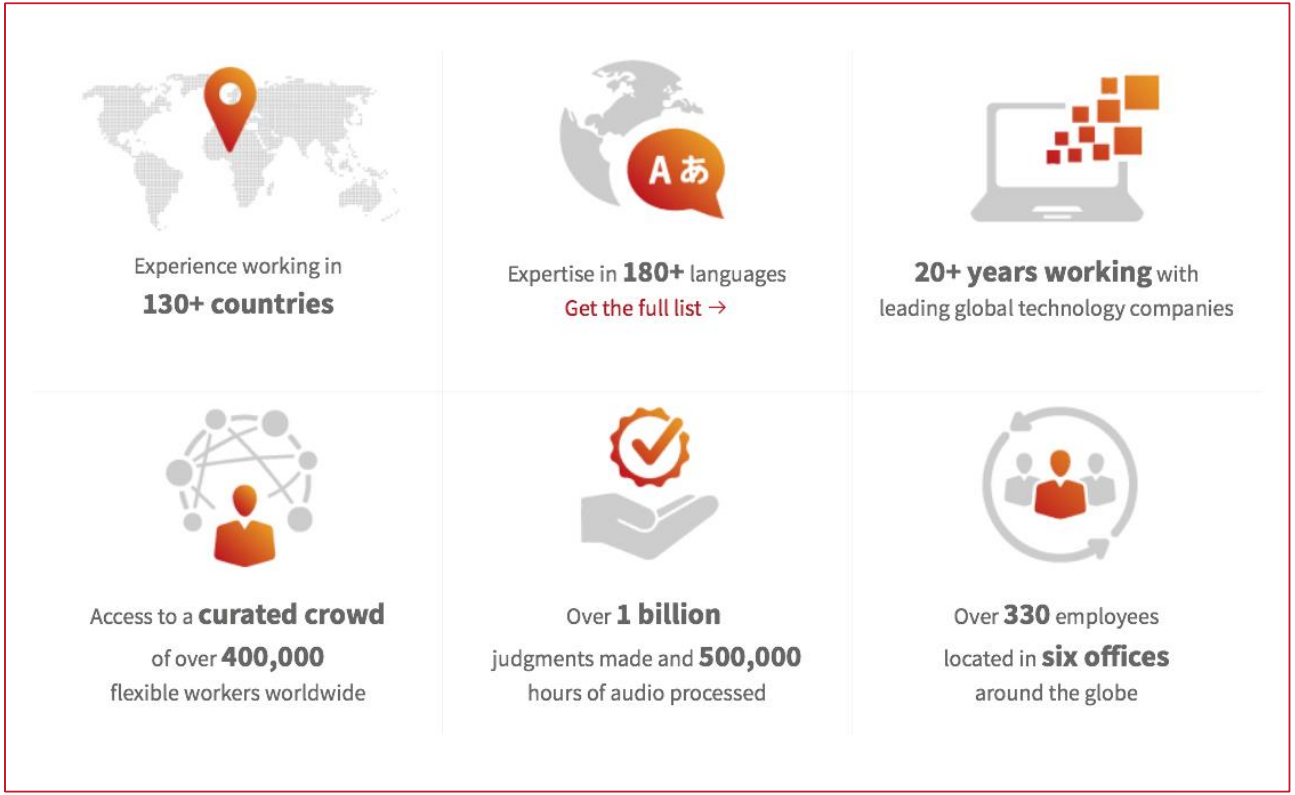
appen™

Got Data?

The Importance of High-Quality Training Data for Building Effective Language-Based Solutions

Appen: Data with a Human Touch



Experience working in
130+ countries

Expertise in **180+** languages
Get the full list →

20+ years working with
leading global technology companies

Access to a **curated crowd**
of over **400,000**
flexible workers worldwide

Over **1 billion**
judgments made and **500,000**
hours of audio processed

Over **330** employees
located in **six offices**
around the globe

The need for annotated data

The role of human-annotated linguistic data in development of AI systems that include NLP components is critical but often unacknowledged.

Where does the data come from?

(graduate students?)

Choice 1: Buy it or Build it?

Licensable Public Data?

Pros:

- Cheap!
- Results easily compared
- Known quality thresholds

Cons:

- Not specific to your domain
- Solves somebody else's problem
- Language coverage

Roll your own?

Pros:

- There's no data like your own data
- Tailor labels to your problem
- In your target (language) market

Cons:

- Must define the label set
- Must find and train annotators
- Quality measures?

Choice 2: What’s my label set?

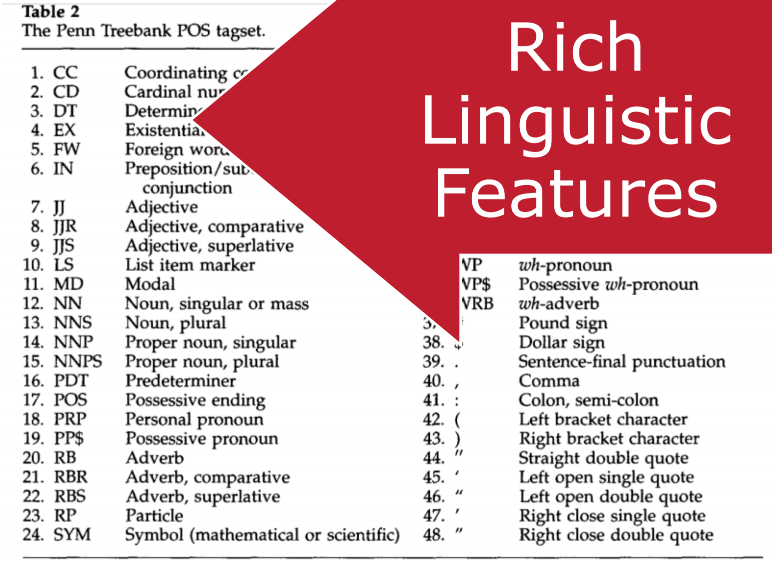
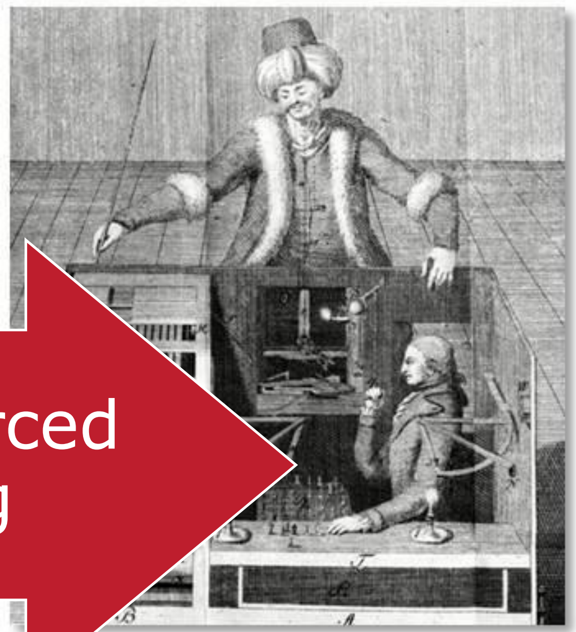


Table 2
The Penn Treebank POS tagset.

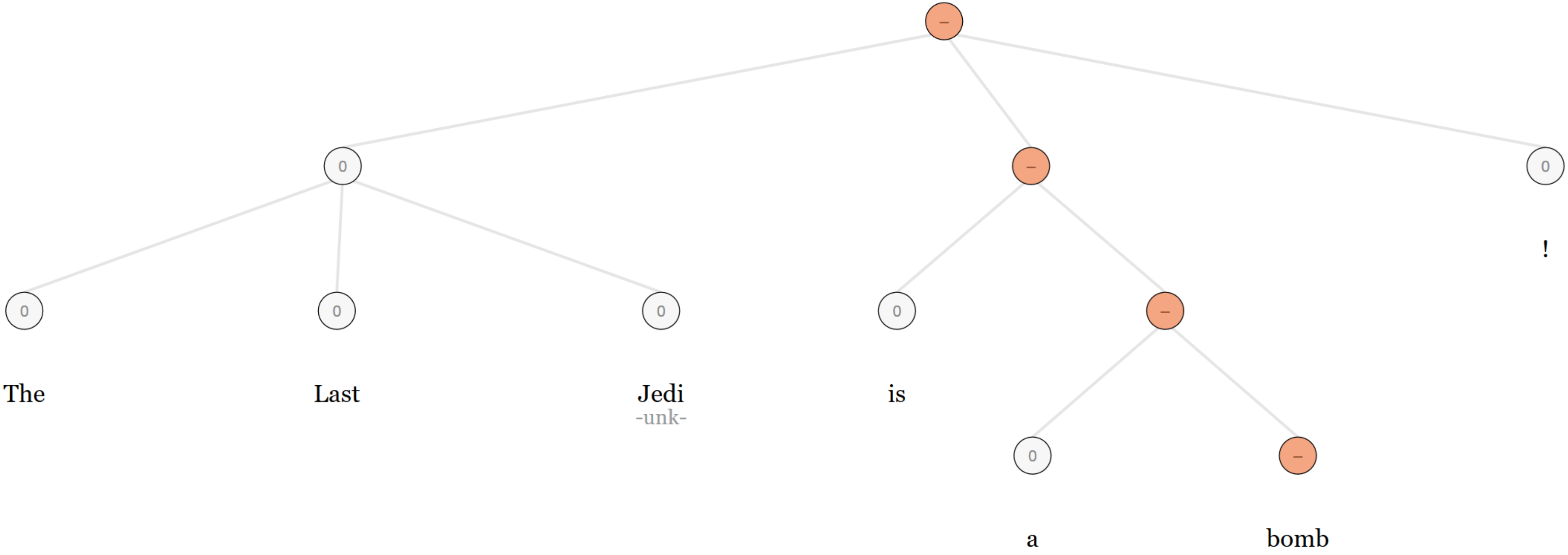
1.	CC	Coordinating c			
2.	CD	Cardinal nu			
3.	DT	Determin			
4.	EX	Existentia			
5.	FW	Foreign word			
6.	IN	Preposition/sub conjunction			
7.	JJ	Adjective			
8.	JJR	Adjective, comparative			
9.	JJS	Adjective, superlative			
10.	LS	List item marker		VP	*wh*-pronoun
11.	MD	Modal		VP$	Possessive *wh*-pronoun
12.	NN	Noun, singular or mass		VRB	*wh*-adverb
13.	NNS	Noun, plural	37.		Pound sign
14.	NNP	Proper noun, singular	38.	$	Dollar sign
15.	NNPS	Proper noun, plural	39.	.	Sentence-final punctuation
16.	PDT	Predeterminer	40.	,	Comma
17.	POS	Possessive ending	41.	:	Colon, semi-colon
18.	PRP	Personal pronoun	42.	(	Left bracket character
19.	PP$	Possessive pronoun	43.	)	Right bracket character
20.	RB	Adverb	44.	"	Straight double quote
21.	RBR	Adverb, comparative	45.	‘	Left open single quote
22.	RBS	Adverb, superlative	46.	“	Left open double quote
23.	RP	Particle	47.	’	Right close single quote
24.	SYM	Symbol (mathematical or scientific)	48.	”	Right close double quote

Rich Linguistic Features

Crowdsourced labeling

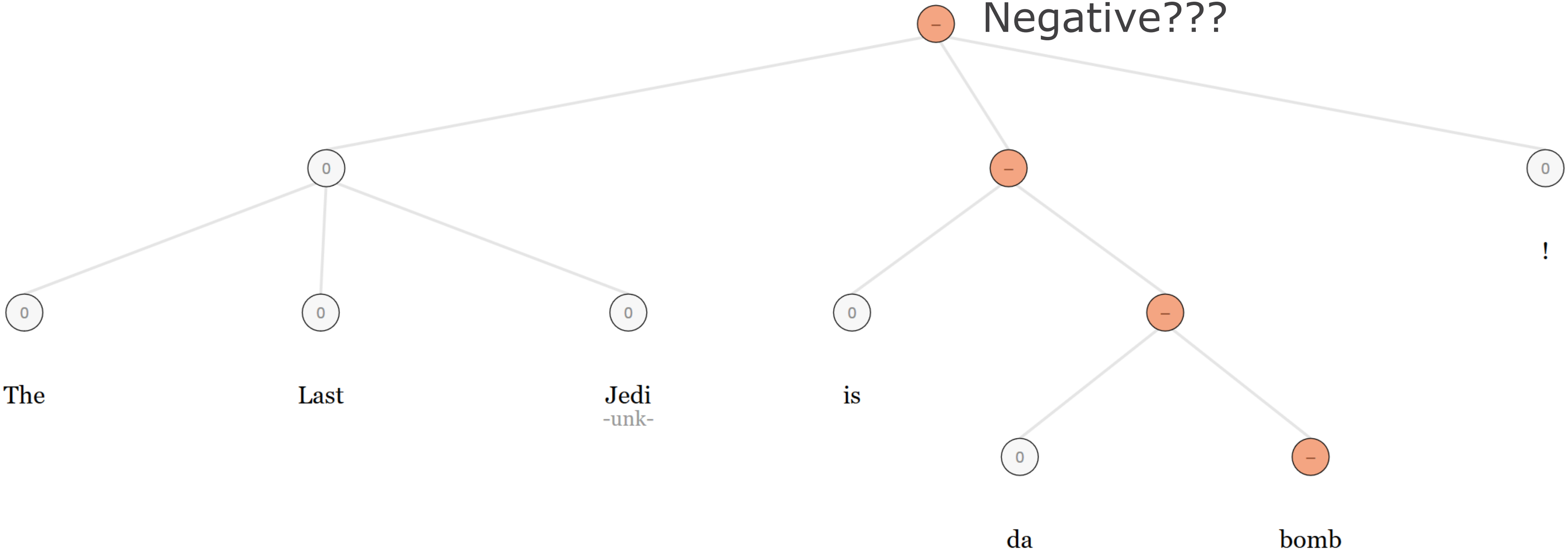


Sentiment Analysis: Movie Review Domain



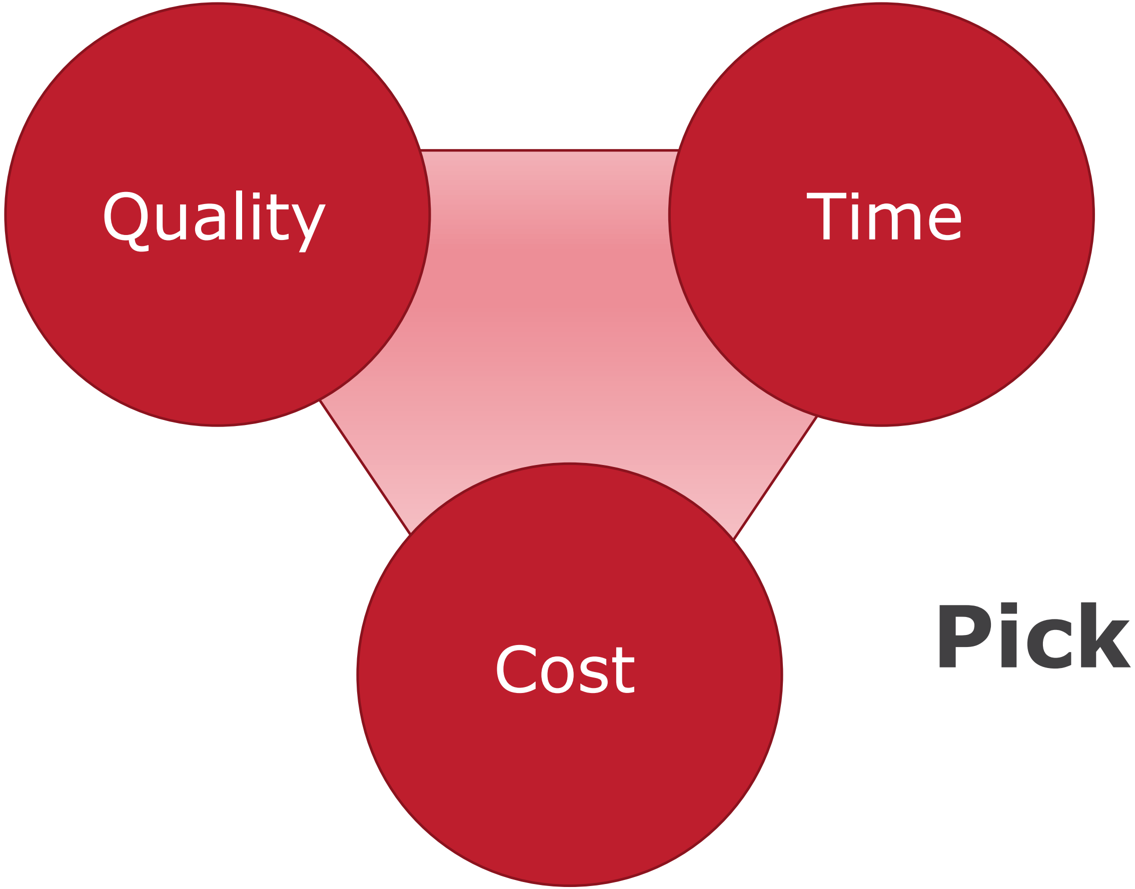
https://engineering.stanford.edu/magazine/article/stanford-algorithm-analyzes-sentence-sentiment-advances-machine-learning

Sentiment Analysis: Movie Review Domain



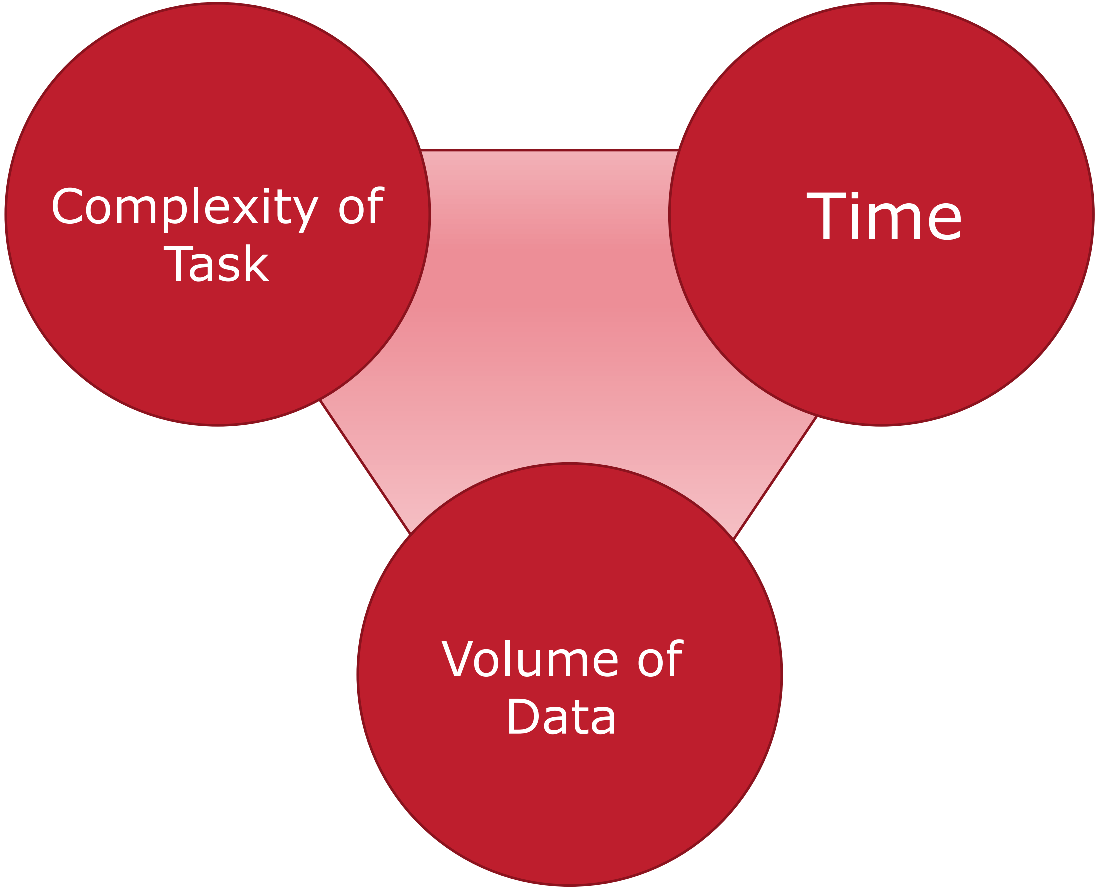
https://engineering.stanford.edu/magazine/article/stanford-algorithm-analyzes-sentence-sentiment-advances-machine-learning

The Iron Triangle …



Pick 2

Choice 3: Complexity or Volume?



Data Annotation LOE (Sentiment)

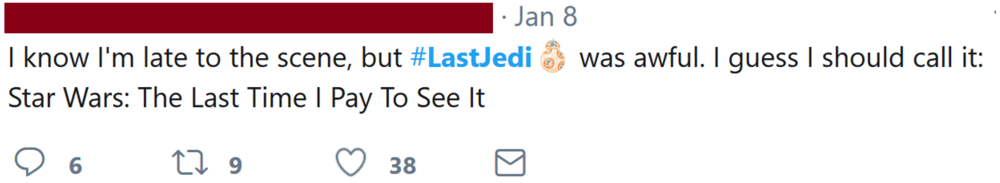
- Documents: Tweets or tweet-like social media text (12-200 chars)
- 3-way judgment at document level (Positive/Negative/Neutral)
- Minimal training requirements (1-2 page instructions)
- Throughput: 250 documents per hour
    - A bit slower depending on the writing system, e.g. 180 per hour for Chinese

→ 10K docs = 40 hours of labor for **1 judgment**

How useful is a single judgment?

Pretty useless, actually …

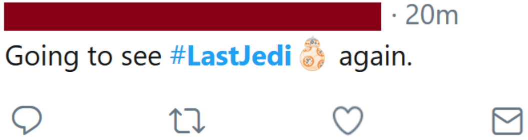
- 30-40% disagreement rate on this task for two minimally trained judges
- 90% consensus with screening and adding a third judge (2 or more judges agree)



[redacted] · Jan 8

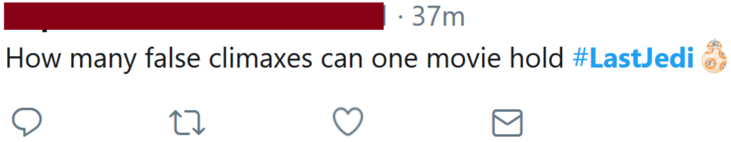
I know I'm late to the scene, but #LastJedi was awful. I guess I should call it: Star Wars: The Last Time I Pay To See It

6 9 38



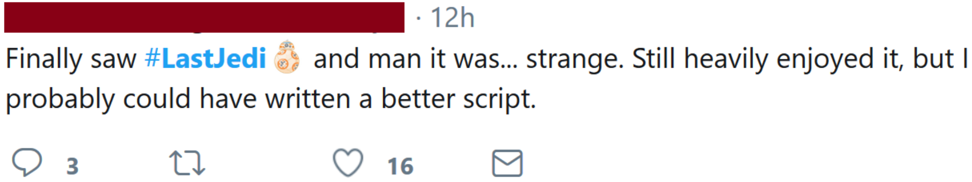
[redacted] · 20m

Going to see #LastJedi again.



[redacted] · 37m

How many false climaxes can one movie hold #LastJedi



[redacted] · 12h

Finally saw #LastJedi and man it was... strange. Still heavily enjoyed it, but I probably could have written a better script.

3 16

Making choices …

- No magic bullets for annotated data creation, only tradeoffs
- Budgeting for data creation is indispensable
- Better results with data that is specific to your domain, annotated to your specifications
- (But be pragmatic about quality targets for newly-designed labeling tasks)
- Small trained/screened crowds with quality feedback loop can yield good results on modest volumes of data

Thank you

James Lyle

Director Custom Linguistic Solutions

jlyle@appen.com