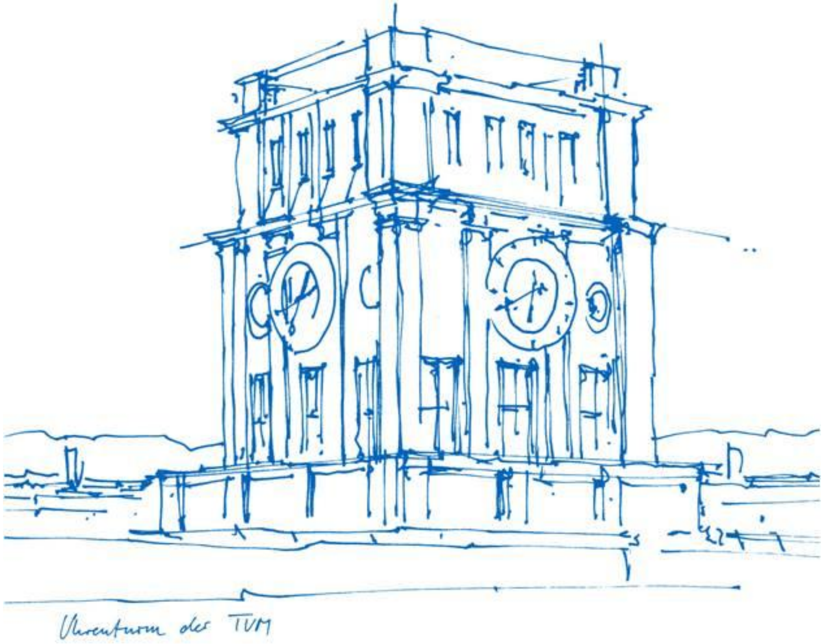
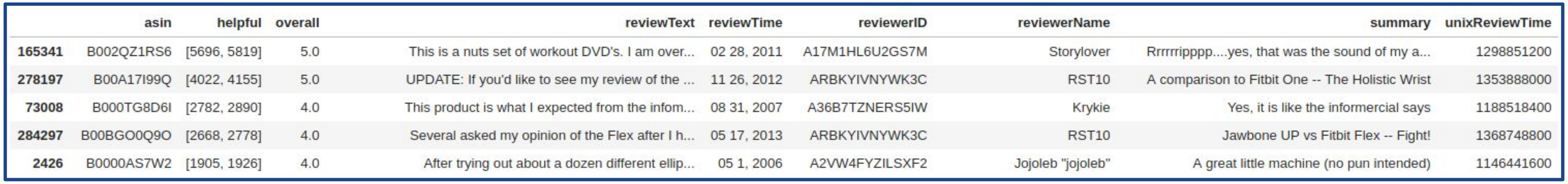
Comparing **Sentiment Analysis** of Discourse Units and Sentences

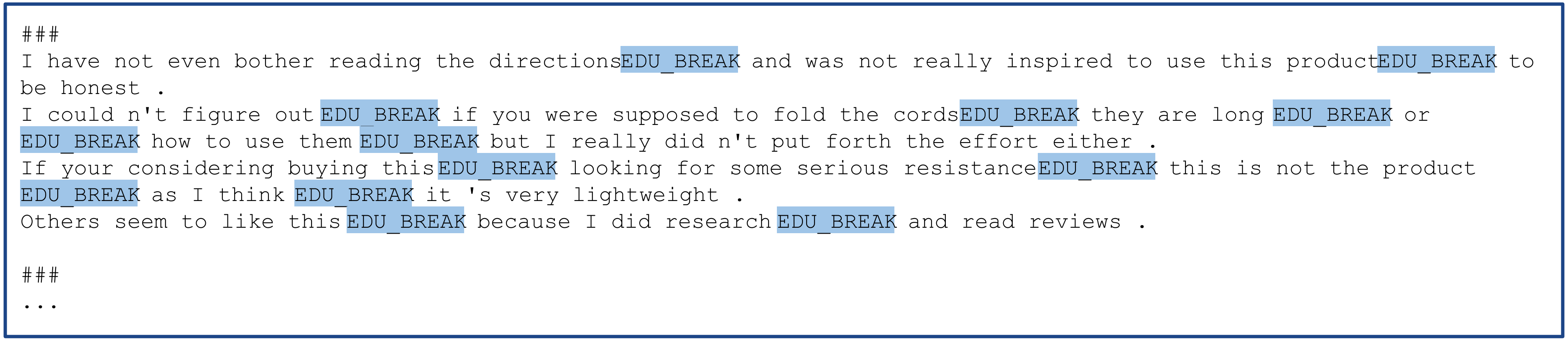
Group 9: Li Canchen, Hendrik Pauthner, Tim Pfeifle
Technische Universität München
Informatics
Research Group Social Computing
Munich, 22. May 2019



Discourse Segmentation



	asin	helpful	overall	reviewText	reviewTime	reviewerID	reviewerName	summary	unixReviewTime
165341	B002QZ1RS6	[5696, 5819]	5.0	This is a nuts set of workout DVD's. I am over...	02 28, 2011	A17M1HL6U2GS7M	Storylover	Rrrrripppp....yes, that was the sound of my a...	1298851200
278197	B00A17I99Q	[4022, 4155]	5.0	UPDATE: If you'd like to see my review of the ...	11 26, 2012	ARBKYIVNYWK3C	RST10	A comparison to Fitbit One -- The Holistic Wrist	1353888000
73008	B000TG8D6I	[2782, 2890]	4.0	This product is what I expected from the infom...	08 31, 2007	A36B7TZNERS5IW	Krykie	Yes, it is like the informercial says	1188518400
284297	B00BGO0Q9O	[2668, 2778]	4.0	Several asked my opinion of the Flex after I h...	05 17, 2013	ARBKYIVNYWK3C	RST10	Jawbone UP vs Fitbit Flex -- Fight!	1368748800
2426	B0000AS7W2	[1905, 1926]	4.0	After trying out about a dozen different ellip...	05 1, 2006	A2VW4FYZILSXF2	Jojoleb "jojoleb"	A great little machine (no pun intended)	1146441600



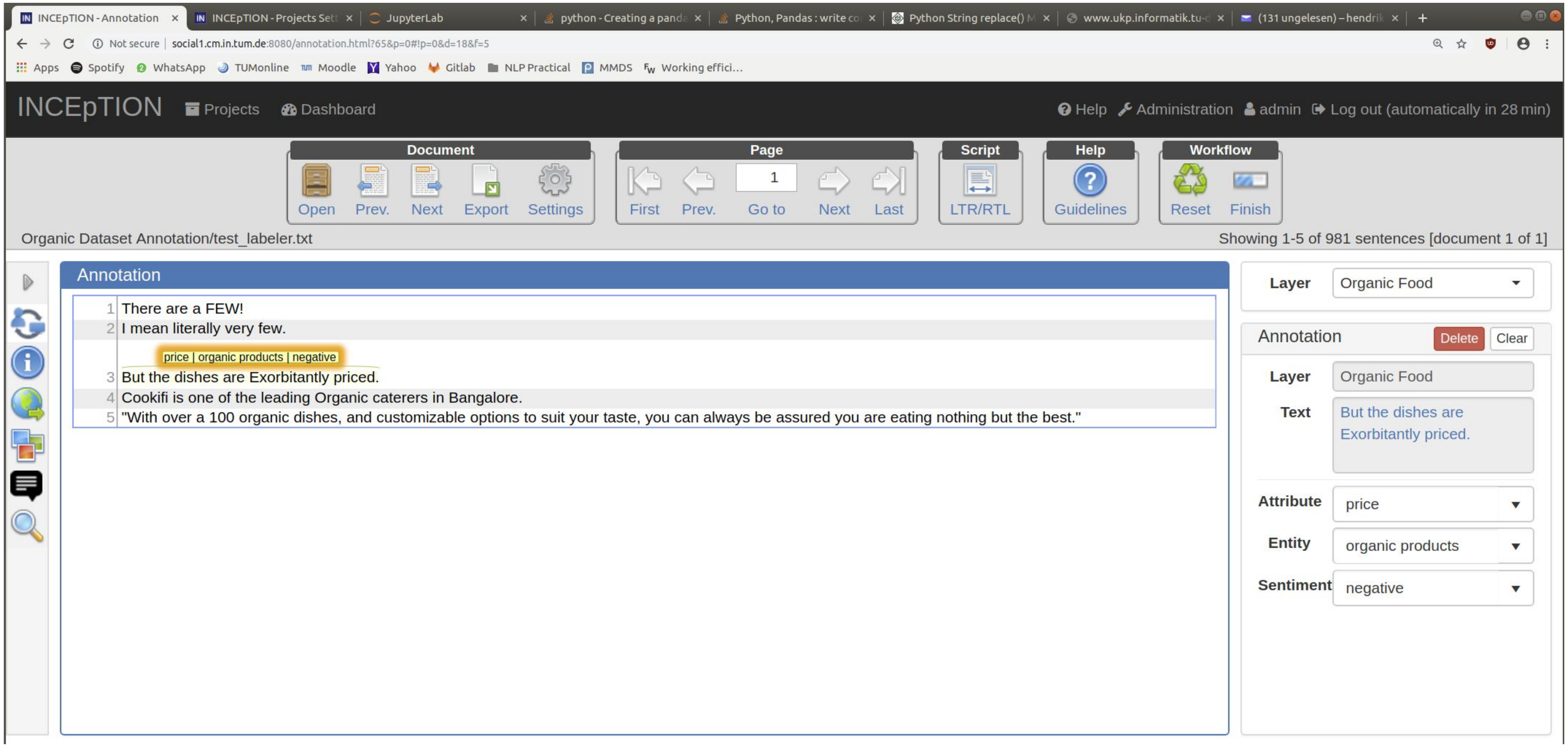
```
###
I have not even bother reading the directionsEDU_BREAK and was not really inspired to use this productEDU_BREAK to
be honest .
I could n't figure out EDU_BREAK if you were supposed to fold the cordsEDU_BREAK they are long EDU_BREAK or
EDU_BREAK how to use them EDU_BREAK but I really did n't put forth the effort either .
If your considering buying this EDU_BREAK looking for some serious resistanceEDU_BREAK this is not the product
EDU_BREAK as I think EDU_BREAK it 's very lightweight .
Others seem to like this EDU_BREAK because I did research EDU_BREAK and read reviews .

###
...
```

We use a discourse parser by Feng and Hirst (2014):

- Yields state-of-the art discourse segmentation results
- Two modes for discourse segmentation: Simple and with a second pass using global features
- Processing 1000 reviews takes **~ 30 min** in simple mode / **~ 45 min** in two pass mode

INCEpTION Annotation Tool



Organic Dataset Annotation/test_labeler.txt

Showing 1-5 of 981 sentences [document 1 of 1]

Annotation

1 There are a FEW!
2 I mean literally very few.
price | organic products | negative
3 But the dishes are Exorbitantly priced.
4 Cookifi is one of the leading Organic caterers in Bangalore.
5 "With over a 100 organic dishes, and customizable options to suit your taste, you can always be assured you are eating nothing but the best."

Layer: Organic Food

Annotation

Layer: Organic Food
Text: But the dishes are Exorbitantly priced.
Attribute: price
Entity: organic products
Sentiment: negative

- Tool for document labeling
- You will be asked to annotate a small subset of the unannotated organic dataset

=> More information coming soon!

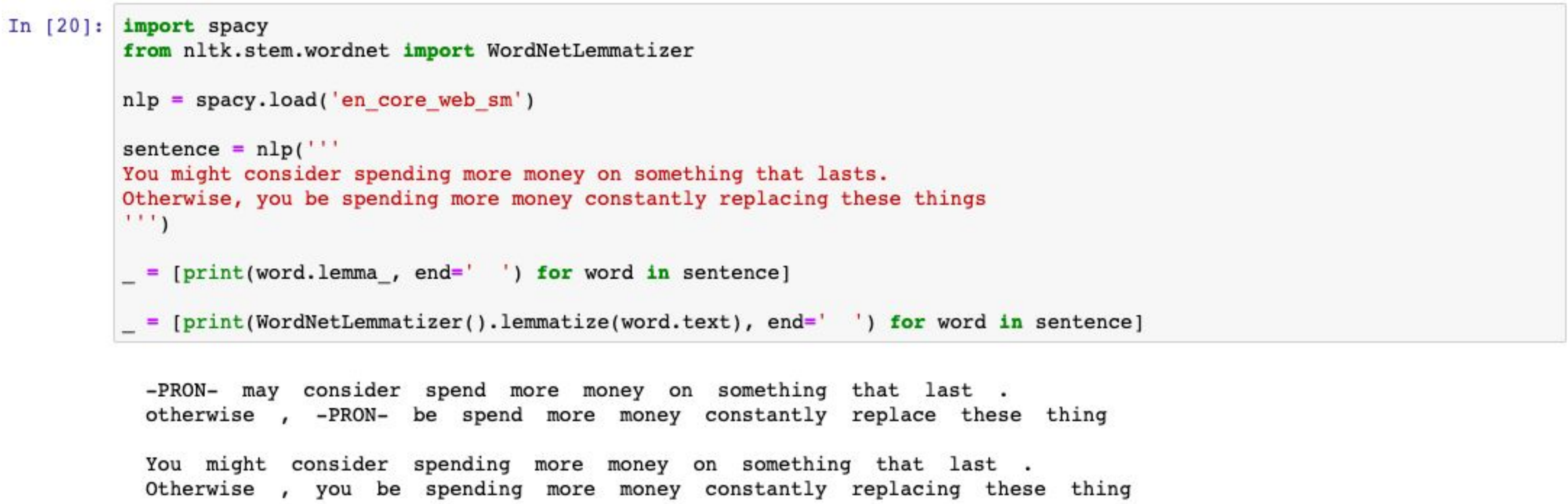
Data Preprocessing

String Cleaning

- Remove numbers, continuous spaces, web links, and separate abbreviations
- Remove stop words and lemmatize

```
The only downside is that there are 3 extra buttons : Power , sleep and wake
EDU_BREAK which it doesn't come with software for .
```

```
['downside', 'extra', 'button', 'power', 'sleep', 'wake', 'n\'t', 'come',
'software']
```



```
In [20]: import spacy
         from nltk.stem.wordnet import WordNetLemmatizer

         nlp = spacy.load('en_core_web_sm')

         sentence = nlp('''
         You might consider spending more money on something that lasts.
         Otherwise, you be spending more money constantly replacing these things
         ''')

         _ = [print(word.lemma_, end='  ') for word in sentence]

         _ = [print(WordNetLemmatizer().lemmatize(word.text), end='  ') for word in sentence]
```

```
  -PRON-  may  consider  spend  more  money  on  something  that  last  .
  otherwise  ,  -PRON-  be  spend  more  money  constantly  replace  these  thing

  You  might  consider  spending  more  money  on  something  that  last  .
  Otherwise  ,  you  be  spending  more  money  constantly  replacing  these  thing
```

Data Preprocessing

Segment Encoding

Word to index mapping

never	1	cell	5
awful	2	phone	6
movie	3	watch	7
better	4	<PAD>	0

Segment encoding

```
['never', 'awful', 'movie'] = [1, 2, 3]
['never', 'better', 'cell', 'phone', 'watch', 'movie'] = [1, 4, 5, 6, 7, 3]
```

Segment Padding

```
[1, 2, 3, 0, 0, 0, 0]
[1, 4, 5, 6, 7, 3, 0]
```

Document Padding

```
[1, 2, 3, 0, 0, 0, 0]
[1, 4, 5, 6, 7, 3, 0]
[0, 0, 0, 0, 0, 0, 0]
```

Data Preprocessing

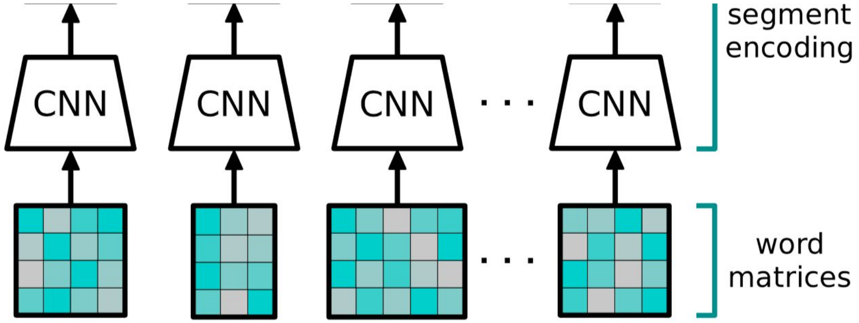
Word Embedding

XLING Feature Mapping

```
[1, 2, 3, 0, 0, 0, 0] -> Lookup Table ->

[[x_1, y_1, …, z_1],
 [x_2, y_2, …, z_2],
 [x_3, y_3, …, z_3],
 [0, 0, 0, …, 0, 0],
 …
]
```

Input to Model



Git + Large Files

Large files, that are modified regularly
→ Problem: Clone takes long (every version of every file downloaded)

Solution:
Git Large File Storage (LFS)

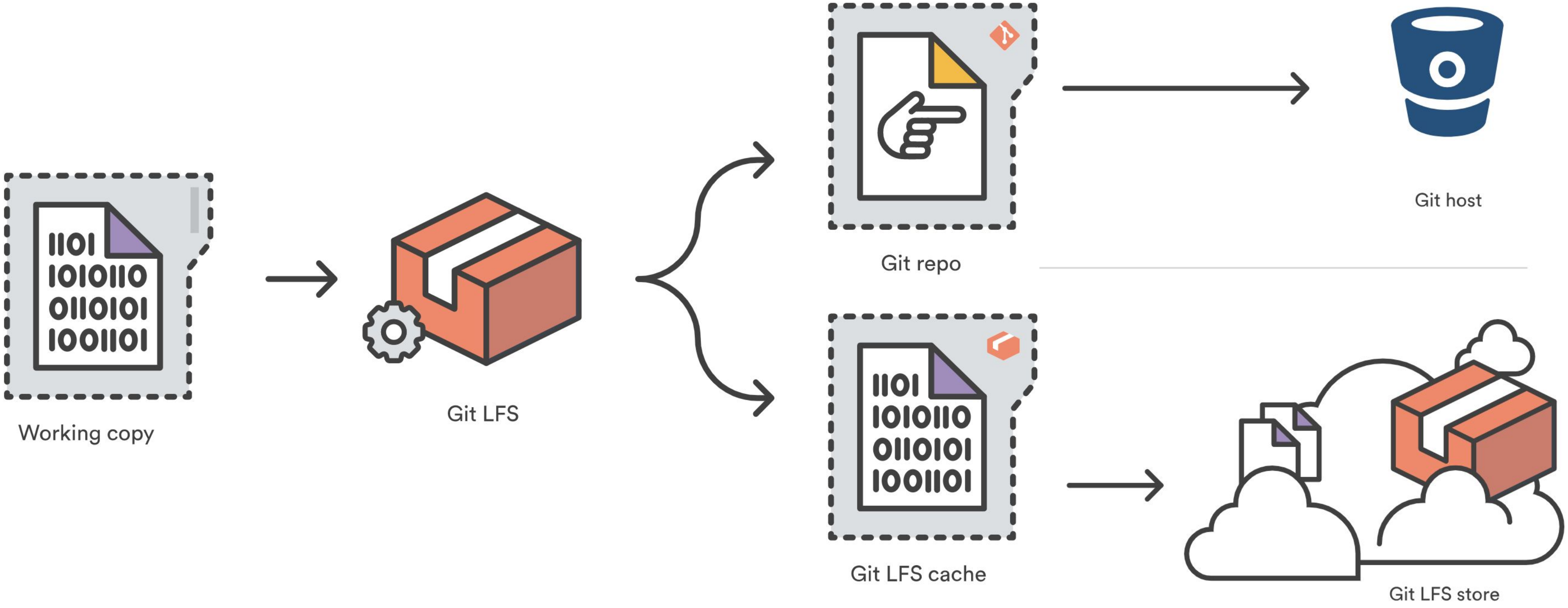
- Download relevant versions **lazily** (during checkout)
- Does not change git workflow

Requirements:

- **Remote:** Git LFS aware host (**Gitlab, ...**)
- **Users:** Git LFS client
- **Repository**: *git lfs install*

Tutorial:
https://github.com/git-lfs/git-lfs/wiki/Tutorial
https://www.atlassian.com/git/tutorials/git-lfs

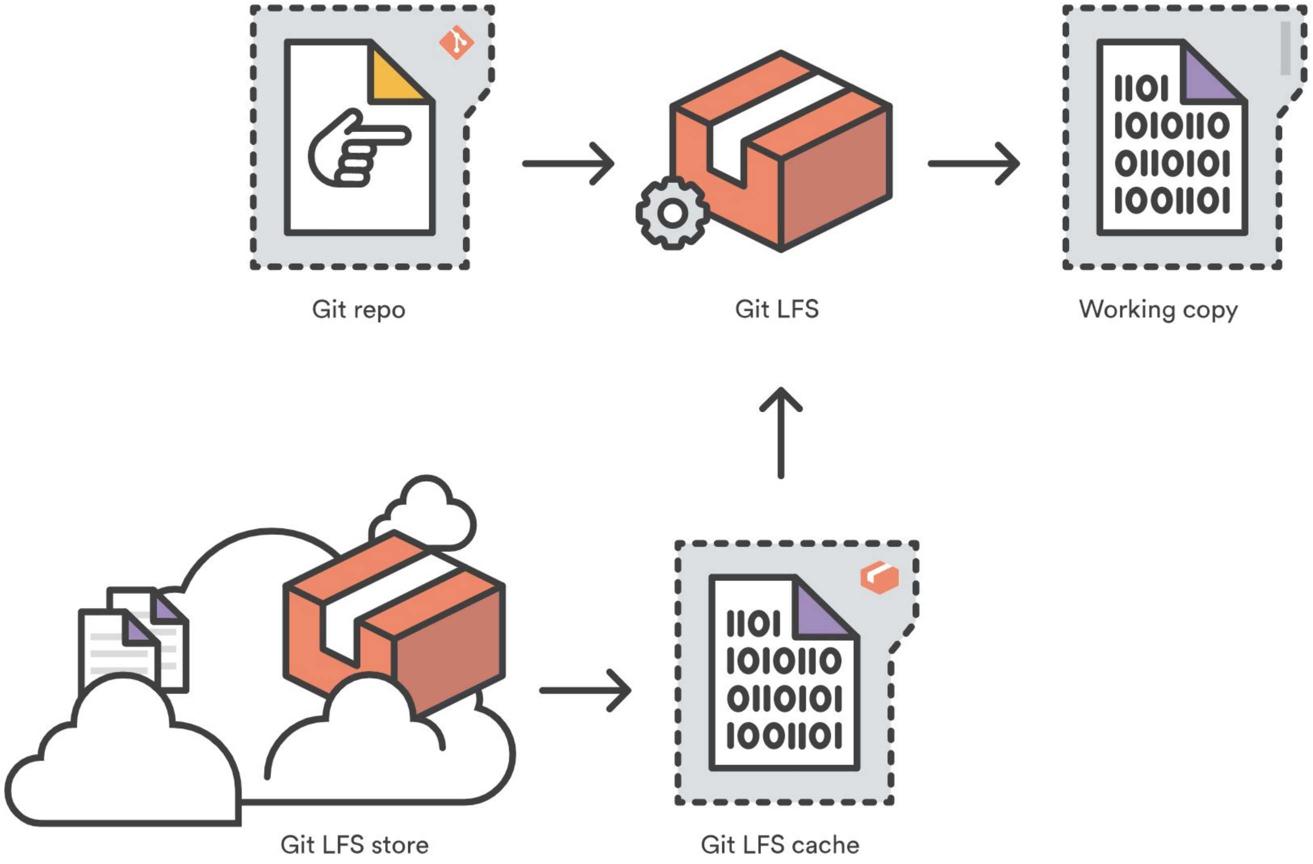
Git LFS: Committing Large Files



git lfs track '*.bin'
→ .gitattributes (check status with ***git lfs ls-files***)

Common git workflow: git add/commit/push/pull

Git LFS: Checkout Large Files



git checkout

Locking support

Next Steps

- Adapt preprocessing for the Amazon review data set
- Run baseline (provided MILNET code in Lua [1])
- Porting MILNET to pytorch
- Creating EDUs from Organic Dataset → Workflow to annotate using Inception

[1] MILNET: https://github.com/stangelid/oposum

Muchas Gracias!

Li Canchen, Hendrik Pauthner, Tim Pfeifle

Munich, 22. May 2019